# Computational Approaches to the Estimation of the Components of Energy Balance in Humans

Ruairi Joseph O'Driscoll

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds
School of Psychology
Faculty of Medicine and Health

March 2021

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4 features a jointly authored publication

- **O'Driscoll, R**., Turicchi, J., Beaulieu, K., Scott, S., Matu, J., Deighton, K., Finlayson, G., & Stubbs, J. (2020). How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. British Journal of Sports Medicine, 54(6), 332–340. https://doi.org/10.1136/bjsports-2018-099643

The candidate took a primary role in designing the research alongside the other co-authors. The candidate also took a primary role in conducting the research by leading the data collection, contributing to data analyses and in writing/editing the manuscript.

Chapter 5 features a jointly authored publication

- **O'Driscoll, R**., Turicchi, J., Hopkins, M., Gibbons, C., Larsen, S. C., Palmeira, A. L., Heitmann, B. L., Horgan, G. W., Finlayson, G., & Stubbs, R. J. (2020). The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours. Health and Technology, 10(3), 637–648. https://doi.org/10.1007/s12553-019-00392-7

The candidate took a primary role in designing the research alongside the other co-authors. The candidate also took a primary role in conducting the research by leading the data collection, contributing to data analyses and in writing/editing the manuscript.

Chapter 6 features a jointly authored publication

- **O'Driscoll, R**., Turicchi, J., Duarte, C., Michalowska, J., Larsen, S. C., Palmeira, A. L., Heitmann, B. L., Horgan, G. W., & Stubbs, R. J. (2020). A novel scaling methodology to reduce the biases associated with missing data from commercial activity monitors. PLoS ONE, 15(6), e0235144. https://doi.org/10.1371/journal.pone.0235144

The candidate played no role in the design of the NoHoW study but worked extensively on data collection. Regarding this publication, the candidate took

a primary role in designing the analysis alongside the other co-authors. The candidate also took a primary role in conducting the research by contributing to data analyses and in writing/editing the manuscript.

Chapter 7 features a publication and another publication currently under review and also another publication:

- **O'Driscoll, R**., Turicchi, J., Hopkins, M., Horgan, G. W., Finlayson, G., & Stubbs, J. R. (2020). Improving energy expenditure estimates from wearable devices: A machine learning approach. Journal of Sports Sciences, 38(13), 1496–1505. https://doi.org/10.1080/02640414.2020.1746088

The candidate took a primary role in designing the research alongside the other co-authors. The candidate also took a primary role in conducting the research by leading the data collection, contributing to data analyses and in writing/editing the manuscript.

## Acknowledgements

This thesis would not have been possible without the hard work, dedication and support of many and before I present this thesis, I must extend my thanks to them.

To my primary supervisor, Professor James Stubbs, who taught me innumerable lessons (both scientific and non-scientific). James' rigorous approach to scientific inquiry can be seen in this thesis and is what I aspire to. To Professor Graham Finlayson and Dr Cristiana Duarte, their calm and supportive approaches helped me through the most challenging years of my life.

To Jake, his frustratingly incisive comments on every chapter in this thesis made for a far stronger PhD.

To Dr Catherine Gibbons, Dr Mark Hopkins, Professor John Blundell and Dr Graham Horgan, for their extensive support and advice throughout and to all members of the human appetite research unit, particularly Fiona, Nuno, Pauline and Kristine. I am grateful to all members of the NoHoW consortium, who worked tirelessly for 5 years to collect the dataset used in this PhD.

To my friends in Leeds and beyond, who offered a refreshing escape from this all-consuming project.

To my parents, Fiona and Denis, for instilling the values of hard work and dedication required for a PhD and for their unwavering support throughout my entire life.

Lastly, to Ella. Words cannot express the kindness, love and support she has shown over the last 4 years. This would not have been possible without her.

## Abstract

**Background:** Continuous, long-term measurement of energy balance behaviours is a significant challenge and of great scientific interest to the field of energy balance and a multitude of related fields. Methodologies such as doubly labelled water (DLW) are infeasible in large scale studies because of their expense. Recent developments in wearable technologies may offer an opportunity to overcome this issue, but uncertainty exists regarding their accuracy. Should accurate estimates of energy expenditure (EE) be obtainable from such devices, it will be possible to incorporate estimates into validated mathematical models to estimate the change in energy intake (EI), in free-living subjects.

**Objectives:** This thesis aimed to examine methods to estimate EE from wearable sensors in free-living subjects participating in the NoHoW trial, a weight loss maintenance intervention.

**Methods:** A series of studies were conducted to investigate the validity of EE estimates from the manufacturer estimates of the Fitbit charge 2, and machine learning models trained on the sensor outputs. Both manufacturer estimates and model predictions were compared in free-living and used to estimate PAEE and $\Delta EI$ in the NoHoW trial.

**Results:** Laboratory validation studies indicated that the manufacturer estimates of the Fitbit charge 2™ were inaccurate and subsequently, that machine learning models could provide more accurate estimates of EE. Comparisons were made to an established research-grade armband, the SenseWear armband mini™ which showed that the manufacturer estimates were in slightly better agreement than the developed algorithms. In the application of several EE estimation methods to the NoHoW dataset, $\Delta EI$ could be estimated and this demonstrated that caloric restriction was greatest in the earlier phases of the intervention and this diminished as time progressed.

**Conclusions:** Digital tracking technologies are providing novel opportunities for physiological research. This thesis took positive steps towards developing a methodological framework for the estimation of free-living EE, which will have implications for energy balance and related fields. Future work will examine the models developed in this thesis against DLW

measurements, and evaluate energy balance modelling in a wider range of subjects and circumstances.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1 – General Introduction

The prevalence of overweight and obesity has increased by three-fold in the last 40 years (Ells et al., 2018) and it has been estimated that by 2050, 60% of males and 50% of females may have obesity (Agha & Agha, 2017; Lobstein, 2007). Overweight and obesity, which are conditions characterised by an accumulation of excess adipose tissue in the body, represent a public health concern owing to the associated comorbidities, which include hypertension, type 2 diabetes, non-alcoholic fatty liver disease, coronary heart disease, stroke, cancer and osteoarthritis (Greenway, 2015; Heymsfield & Wadden, 2017). Obesity is associated with an increase in all-cause mortality (Flegal et al., 2013), impaired quality of life (Taylor et al., 2013) and substantial economic burden, estimated to be near 2.8% of the global gross domestic product (Tremmel et al., 2017). The personal, health and economic implications of the obesity epidemic are monumental.

A bodyweight loss of 5% is considered the minimum weight loss required to produce a clinically significant improvement in metabolic health outcomes and is, therefore, a common goal in weight loss interventions (Magkos et al., 2016). Although approximately 40% of adults are attempting to reduce their body weight in the western world (Santos et al., 2017), such attempts are typically unsuccessful in the long term as recidivism to baseline weight or beyond is common (Wing & Phelan, 2005). Indeed, less than 20% of individuals are successful in maintaining a 10% body weight loss for a year or more (Kraschnewski et al., 2010). A weight regain of 2-6% is associated with a return to baseline in the health markers that initially improved with weight loss (Swift et al., 2018). In this sense, the prevention of weight regain after weight loss may be considered the most pressing problem in obesity therapeutics (MacLean et al., 2015). Though that is not to undermine the importance of other factors, namely weight loss and the prevention of weight gain.

The regain of body weight after weight loss is the product of a prolonged positive energy balance, as dictated by the laws of thermodynamics (Bray & Bouchard, 2020). Despite this ineluctable truth, viewing weight change in such terms exclusively ignores how genetics, psychology, physiology and environmental factors interact to determine energy balance and body weight (Butland et al., 2007; MacLean et al., 2015). Our understanding of the

mechanisms by which energy balance is achieved and weight regain is prevented is yet to be fully elucidated.

## 1.1 Energy balance

According to the first law of thermodynamics, energy is neither created nor destroyed, it is converted between forms. In human physiology, this suggests that any perturbation of energy stores (ES) corresponds to the difference between energy intake (EI) and energy expenditure (EE) (Hill et al., 2012):

$$ES = EI - EE$$

The rate of EI is determined by the ingestion of dietary macronutrients and the rate of EE through immediate heat loss or via chemical, mechanical or electrical work. If the sum of energy contained within the body is consistent over a sustained period (i.e. $\Delta$ ES $\approx$ 0), EI must be in equilibrium with EE, a state referred to as energy balance. If EI is consistently greater than EE (i.e. positive energy balance), energy will be stored. In a negative energy balance, EI is less than EE, resulting in a net loss of energy from the body.

### 1.1.1 Energy intake

Humans energy intake is the sum of dietary macronutrient intake: carbohydrate, protein, and fat and whilst not a macronutrient, alcohol intake contributes a variable amount to EI. The term ‘gross energy’ refers to the total chemical energy within a particular food and is distinct from ‘metabolisable energy’ which refers to the difference between the gross energy and the losses and therefore denotes the amount of energy available for biological processes. These losses are via faecal and combustible gases, urinary loses and body surfaces heat losses (Elia & Cummings, 2007). Digestibility of foodstuffs influences the metabolisable energy and varies markedly between foods depending on several factors including the physical structure and its fibre content, which hinder the access of digestive enzymes (Hall et al., 2012). Commonly used metabolisable energy coefficients for the macronutrients are carbohydrate ~4 kcal/g, protein ~4 kcal/g, fat ~9 kcal/g and alcohol ~7 kcal/g though these represent rounded population averages and can vary considerably between individuals (Hall et al., 2012) and nutrient subtypes (Elia & Livesey, 1992). These

macronutrients contribute to the production of energy directly from their dietary forms or are mobilised from their stored forms, to enter energetic pathways. Glucose from dietary carbohydrate provides energy via glycolysis, which produces two adenosine-triphosphate (ATP) anaerobically and pyruvate molecules. Pyruvate is then converted to acetyl coenzyme A, which enters the metabolic pathway known as the Krebs cycle alongside fatty acids. The cycles produce the cofactors NADH and $FADH_2$, which shuttle electrons in the electron transport chain (Medeiros and Wildman, 2018, pp 13-17). The transfer of electrons to oxygen in the electron transfer chain provides 34 ATP molecules (Cooper, 2018. pp 81- 85).

### 1.1.2 Energy expenditure

Before discussing the components of EE at a high level, it is important to define precisely why and how energy is expended from a thermodynamic and bioenergetic perspective. Almost all cellular activity requires energy to proceed. The second law of thermodynamics states that entropy (i.e. disorder or randomness) will increase within a closed system (Dulloo, 2010). However, many cellular processes do not appear to adhere to this law and move towards a more orderly, low entropy state. This contradiction is explained by the open (i.e. not isolated) nature of cells; they can dissipate heat to the external environment, such that entropy can decrease. Thermodynamically, the change in Gibbs free energy ($\Delta G$) of a reaction combines enthalpy (i.e. heat, $\Delta H$), entropy ($\Delta S$) and temperature ($T$) and is formulated as $\Delta G = \Delta H - T\Delta S$. Biological activity proceeds toward minimising free energy (i.e. $\Delta G < 0$) though many reactions themselves are thermodynamically unfavourable in nature (i.e. $\Delta G > 0$). To permit biological reactions in the necessary direction, a coupled energetically favourable reaction is required (Cooper, 2018. pp 81- 85). This role is fulfilled by the free energy-storing molecule, ATP. This molecule, which is comprised of an adenine base, a ribose and three anhydride-bound phosphates provides energy by the hydrolysis of anhydride bonds to yield adenosine-diphosphate (ADP) and $\Delta G = -7.3$ kilocalories/mol. The second hydrolysis of ADP provides adenosine-monophosphate (AMP) and $\Delta G = -3.4$ kilocalories/mol. Though favourable cellular conditions mean that the $\Delta G$ is closer to –12 kcal/mol for the hydrolysis of ATP (Medeiros and Wildman, 2018, pp 13-17). The free energy provided by these reactions provides the essential energy for reactions involved in membrane transport, molecular synthesis and mechanical work.

At a higher level, EE may be divided into three distinct components; resting metabolic rate (RMR), dietary-induced thermogenesis (DIT) and activity energy expenditure (AEE) and the latter can be further subdivided into the EE during physical activity (PAEE) and non-exercise activity thermogenesis (NEAT). The summation of these components over 24 hours represents the total daily EE (TDEE) of a biological system (Ravussin et al., 1986). Figure 1.1 provides a simulated illustration of how variance in these components determines TDEE, which is informed by a previous illustration (Frayn & Evans, 2019. pp 335).

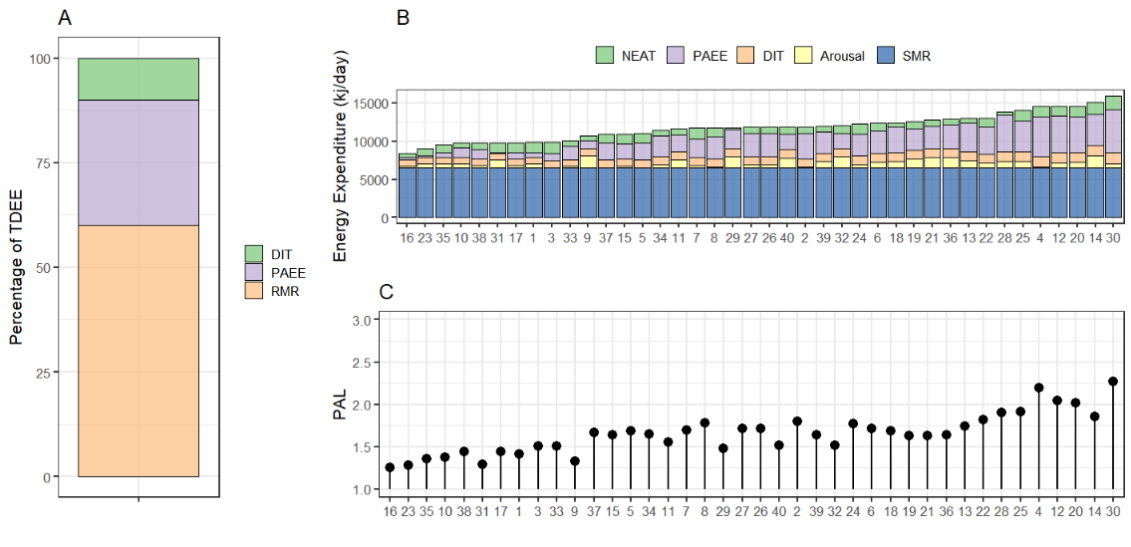



**Figure 1.1** A simulated dataset to illustrate how the components of TDEE can vary in a group of subjects.

A) Typical distribution of DIT, PAEE and RMR.

B) A simulation providing an example of how TDEE can vary depending upon their constituents. The SMR is held constant at 6500 KJ/day and DIT represents 10% of TDEE. The RMR is the sum of arousal and SMR per day. The simulation allowed arousal, PAEE and NEAT to vary. The x-axis represents each of the simulated subjects, ordered by TDEE.

C) The physical activity level (PAL) is calculated as TDEE/RMR. The x-axis represents each of the simulated subjects, ordered by TDEE. Abbreviations: Dietary induced thermogenesis (DIT), non-exercise activity thermogenesis (NEAT), physical activity energy expenditure (PAEE), Resting metabolic rate (RMR), sleeping metabolic rate (SMR), total daily energy expenditure (TDEE).

#### 1.1.2.1 Resting metabolic rate

In a typical western adult, RMR represents the largest component of TDEE, comprising 60-75% of TDEE (Lam & Ravussin, 2016). It refers to the energetic cost to maintain physiological function whilst at rest (Westerterp,

2000). The RMR is measured under resting conditions when a subject is awake, supine, fasted and free from any physical or mental arousal, to ensure the measured value represents the sum of the metabolic cost of essential functions in the resting state (McArdle et al., 2010. pp 191-202). An important distinction is made between RMR and metabolic rate during sleeping; the latter can be up to 20% less than the metabolic rate during waking, explained primarily by the effects of arousal (Dulloo, 2010), though the average sleeping metabolic rate is ~95% of the RMR (Goldberg et al., 1988). Approximately 70%-80% of individual variation in RMR is attributable to the metabolic rate of fat-free mass (FFM), as FFM is comprised of metabolically active organs such as the liver, brain, kidney and skeletal muscle mass (Dulloo, 2010). In a healthy reference adult, fat mass (FM) contributes minimally to overall metabolic rate (Wang et al., 2011) and further variance is attributable to sex, age, ethnicity, metabolic adaption and environmental factors (Bogardus et al., 1986; Fothergill et al., 2016; Weyer et al., 1999).

#### 1.1.2.2 Activity energy expenditure

To perform spontaneous or voluntary exercise, muscular work is required. The absolute energy requirement of a particular activity varies depending on the duration, intensity and body mass of the subject and therefore AEE is often normalised relative to RMR. The most intense activities demand energy costs of >14 x RMR (Ainsworth et al., 2011; Durnin, 1991), though world record standard performances are likely to be higher, depending on the duration of the event (Thurber et al., 2019). Variability in behaviour and lifestyle of individuals means that the physical activity level (i.e. TDEE/RMR) can range from 1.2 to 5, or even higher in the most extreme examples (Black et al., 1996; Westerterp, 2001). Some additional variance may be attributable to exercise economy, a phenomenon whereby the EE for a given volume of physical activity is reduced and this is a common adaption to exercise training (Hunter et al., 2015). NEAT, which is considered by some as distinct from AEE, refers to EE through non-volitional movements such as fidgeting or postural control (Lam & Ravussin, 2017) although the definition of NEAT remains inconsistent, with some researchers including daily tasks such as walking (Garland et al., 2011). Recent work has compiled a body of data collected during endurance events to demonstrate the relationship between duration and physical activity level (PAL) = EE/RMR (Thurber et al., 2019). The parameters from the logarithmic regression reported in this work are plotted in unlogged form in figure 1.2. Though the contributing data to

these models is not extensive, a clear curvilinear pattern emerges, which is to say that the potential intensity with which an activity can be performed decreases with increasing duration.

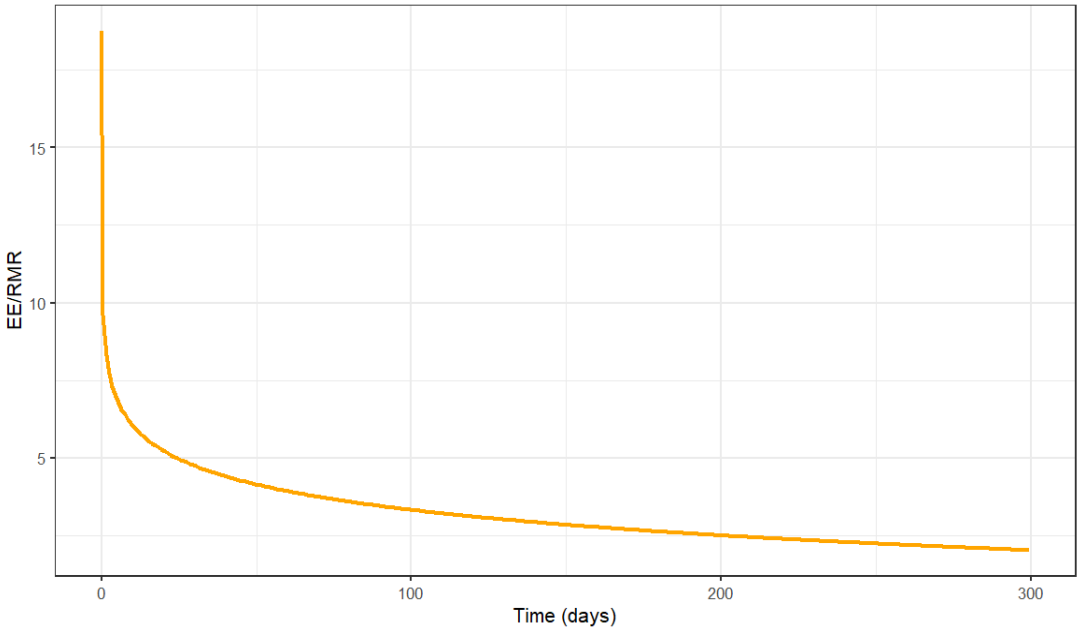



**Figure 1.2** Predicted maximal EE/RMR relationship for epochs ranging from minutes to 300 days.

EE/Basal metabolic rate (BMR) values are obtained by separate models of < 0.1 days and 0.5-300 days (see Thurber et al., 2019).

### 1.1.2.3 Dietary induced thermogenesis

Increases in EE above fasting levels following the ingestion of a meal is referred to as DIT. This is the obligatory energetic costs associated with the gastrointestinal tract activity, conversion of gross energy to metabolisable energy and the energy cost associated with storing foods. For example, the formation of glycogen for storage from glucose is an energy-consuming process, requiring the hydrolysis of ATP and uridine-triphosphate (Frayn and Evans, 2019. pp 335). This is a challenging process to measure accurately (Ruddick-Collins et al., 2013) although it is generally proportional to the total EI, thus it is often approximated at 10% of the TDEE (Westerterp et al., 2004), assuming that TDEE ≈ EI. Despite this approximation, DIT varies based on the macronutrient composition of the meal and can be increased by exercise training (Byrne & Hills, 2018), proportionately reduced in people with obesity compared to lean subjects (de Jonge & Bray, 1997) and increases in response to extreme overfeeding (Pasquet et al., 1992).

### 1.1.3 Energy storage

A change in ES (i.e. $\Delta ES \neq 0$) represents the net change of the dietary macronutrients stored within the body. Energy imbalances are accounted for by deposition or metabolism of body fat, protein, and glycogen and change in mass is determined by the tissues comprising the loss or gain. Energy densities of glycogen, protein and fat approximate to 17.6, 19.7, and 39.5 MJ/kg, respectively (Livesey & Elia, 1988). Carbohydrate is stored as intracellular glycogen in the liver and skeletal muscle and despite carbohydrate being the primary provider of dietary energy in typical western diets, the total amount of stored glycogen is comparatively small with approximately 400 grams in skeletal muscle and 100 grams in the liver (McArdle et al., 2010. pp 13-15). In non-diabetic subjects, blood glucose is extremely tightly maintained and glycogen fluctuates markedly to facilitate this (Galgani & Ravussin, 2008). Each gram of glycogen is associated with ~3 grams of water and this highlights one mechanism by which short-term body weight fluctuations may not reflect true changes in ES (Bhutani et al., 2017). The storage of protein represents approximately 30% of the ES of an adult man however this value is subject to change in response to weight gain and non-dietary stimuli (i.e. resistance training) (Galgani & Ravussin, 2008). As with carbohydrate, protein is associated with water and is therefore not a particularly energy-dense storage medium. Most proteins serve essential biological functions and thus, protein balance is very tightly maintained in humans when at or close to energy balance (Abbott et al., 1988). Lipid is stored as triglycerides in adipocytes and represents the largest potential energy store in humans. Adipose tissue functions as the primary ‘energy buffer’ for the body and is used as the primary deposit for long-term energy imbalances (Abbott et al., 1988). The mechanism for increasing the ES in fat is through increases in the number and capacity of adipocytes and adults with morbid obesity may have 4 times the adipocytes as lean adults and twice the triglyceride per adipocyte (Hirsch & Knittle, 1970). Importantly, the loss of stored energy from the body is largely accounted for by a change in the size rather than the number of adipocytes (Maclean et al., 2015).

Assuming a typical western diet being consumed in a man weighing 70kg ~10.5 MJ/day EI is required to maintain energy balance, although this value has substantial scope to vary with physical activity. Energy obtained from dietary fat and carbohydrate is ~4.2 MJ/day and a further 2.1 MJ/day is obtained from dietary protein. The storage components of carbohydrate are ~8 MJ, protein is ~170 MJ and fat is ~525 MJ (Lam & Ravussin, 2017). The

proportion of the stored energy oxidised daily for carbohydrate, protein and fat is ~50%, ~1.3% and <1%, respectively. In a positive energy balance, the oxidation rates (as a proportion of stored energy) of carbohydrate and protein increase, reflecting a tighter homeostatic regulation of these stores. By contrast, the oxidation of fat is relatively constant whilst the adipose mass expands, manifesting in weight gain (Lam & Ravussin, 2017).

### 1.1.4 Interactions between the components of energy balance

Energy balance is a dynamic system and therefore any behaviour which alters one component of the energy balance equation cannot be simply considered as an additive change to ES (Hall, 2008). Indeed, the magnitude of an energy deficit or surfeit decreases as energy is accumulated or metabolised as a fuel because the mass of the subject changes (Melby et al., 2017). Longitudinal changes in both FM and FFM are determined by initial ES, body composition, the magnitude of the energy imbalance and the physical activity status of the subject (Forbes, 2000). During negative energy balances, both behavioural and physiological adaptions occur, with TDEE decreasing by up to 25% after 10% body weight loss (Rosenbaum & Leibel, 2010). The marked reduction may exceed what would be predicted based on body composition changes (Leibel et al., 1995) and this phenomenon is termed adaptive thermogenesis (Müller & Bosy-Westphal, 2013). Adaptive thermogenesis is thought to be attributable to decreases in NEAT, increases in skeletal muscle efficiency and decreases in RMR (Rosenbaum & Leibel, 2010), although much of the research is equivocal and substantial inter-individual differences exist (Melby et al., 2017). Adaptive responses to changes in weight are asymmetric in their intensity, favouring weight regain (Müller et al., 2010). Notable investigations of the adaptive and behavioural responses to weight loss include research conducted on the participants in the ‘Biggest Loser’ television program, in which subjects lost ~60kg from a baseline weight of ~150kg, followed by a subsequent regain of ~40kg at 6 years. After this regain period, RMR was suppressed at ~500 kcal below predicted values based on body composition, implying substantial adaptation to RMR (Fothergill et al., 2016). Second, the CALERIE trial required subjects to reduce their caloric intake by 25% and showed that this was accompanied by reductions in activity related EE (Martin et al., 2011) and this is probably in addition to adaptive changes in body composition and the function of tissues (Stubbs & Turicchi, 2021). Indeed, it may be that conscious dietary restraint and careful attention to lifestyle choices are necessary to

counterbalance these adaptive responses to weight loss (Marlatt et al., 2017).

Positive energy balance appears to elicit smaller physiological compensatory responses to prevent further weight gain, when compared to weight loss. These responses may include a small increase in NEAT (Levine et al., 1999), as well as DIT associated with increased EI and the energetic cost of tissue synthesis (Westerterp, 2013). This asymmetry in energy balance regulation has been illustrated in a recent review of the literature, in which Bray and Bouchard show a strong linear association ($R^2 = 0.88$) between the energy overfed and the $\Delta ES$ (Bray & Bouchard, 2020).

When a subject is at or close to energy balance, the relationship between EI and EE on a daily basis is variable, however it becomes increasingly balanced as the timeframe in consideration increases. Edholm showed that there is no association between EE and EI on a given day but as this measurement period extends to two weeks, the relationship becomes far stronger (Edholm et al., 1955, 1970). This seminal work implies that energy imbalance is tolerable over short time frames but less so as time increases. To capture the fundamental importance of time in the understanding of energy balance, the static energy balance equation can be modified (Alpert, 1990):

$$Rate\ of\ ES = Rate\ of\ EI - Rate\ of\ EE$$

This subtle but important adaption to the energy balance equation encompasses the importance of the time domain in changing ES (Galgani & Ravussin, 2008). Relatively short-term interventions in which EE and EI are perturbed by exercise and dietary energy density to create energy imbalances of 5–6 MJ/day result in compensation to EI and EE of 0.2 and 0.35 MJ/day, respectively (Stubbs et al., 2004). If extrapolated linearly beyond the data, up to 4 weeks would be taken to compensate fully. Subsequent investigations confirm that an exercise-induced negative energy balance compensation occurs over two weeks with clear evidence of inter-subject differences, which mechanistically remain unexplained (Whybrow et al., 2008).

## 1.2 Measures of energy intake

Prolonged energy imbalance is the determinant of a systematic change in body weight and therefore quantifying the components of energy balance has been the subject of scientific research since the 18th century, when Lavoisier and Seguin pioneered the methodology of indirect calorimetry. Today, in both laboratory and free-living settings, it is feasible to derive accurate and precise estimates of EE, ES and EI (Hills et al., 2014; Lam & Ravussin, 2016). However, longitudinal research in weight management is limited by the lack of scalable tools to accurately quantify both EI and EE in free-living human subjects, over months and years. The lack of such measurement tools limits quantitative understanding of the extent and mechanisms by which weight maintenance is achieved and contributes to uncertainties and debate regarding the aetiology of human obesity (Williams and Frühbeck, 2009. pp 187 - 208). The remainder of this review provides a critical, but non-exhaustive overview of methodologies currently used in energy balance research. Tables 1.1 and 1.2 summarise methodologies for the assessment of energy or food intake and EE/physical activity, respectively. A distinction is made between self-report measures and more objective approaches throughout because the error associated with self-report measures is often large and unpredictable. When self-report EI and EE estimates are used together, the errors can compound, increasing the probability of erroneous conclusions. This bias creates serious concerns over potentially misguided national health policy. Indeed, prominent researchers have argued that these assessment tools are so inaccurate that they are inappropriate for use in scientific research (Dhurandhar et al., 2015). This issue, as well as potential solutions, are discussed herein.

### 1.2.1 Quantifying energy intake with self-report measures

#### 1.2.1.1 Measurement tools

Self-report measures of EI are ubiquitous within medical and health research settings. These tools include a broad range of assessments which require input from a participant in an attempt to understand their food intake or eating behaviour. These measures vary significantly in their ease of dissemination, participant burden, required researcher expertise and accuracy of EI estimates. Traditional assessment techniques include food diaries, in which subjects record the food and beverages consumed over a period (typically up to 7 days) and food recalls, in which a trained researcher interviews a subject about their previous consumption. With these

assessments and the many varieties of self-report tools, the quality of the assessment is highly dependent on the technical skill of the experimenters and the recall of the subjects (Lam & Ravussin, 2016). More recently, food photography methods have emerged. These techniques require a subject to provide photographs of a meal, which is then analysed relative to known portion sizes by a semi-automated process. The real-time nature of this method allows for reminders and prompts to be sent by researchers, which in theory should minimise recall bias. Food photography has shown promise in small studies relative to a doubly labelled water (DLW) criterion, with a mean underestimate in EI of <10% when customised prompts delivered at meal times are used (Martin *et al.*, 2012), and this could be related to the elimination of biases associated with retrospective recall. These methods only capture small periods of an individual's diet (Johnson, 2002) and evidence suggests that participants may alter intake behaviour over a measurement period (Trabulsi & Schoeller, 2001), both of these issues create uncertainty around whether these tools are representative of the subject's typical diet.

#### 1.2.1.2 Misreporting of energy intake

Dietary misreporting has been recognised for at least 30 years (Lissner et al., 1989). Misreporting is a phenomenon in which reported EI deviates from the true EI, and this is generally in the direction of underreporting (Murakami & Livingstone, 2015). Two processes are likely to be at play in the degree of misreporting. First, the 'reporting effect' is characterised by an incomplete recording of intake and the 'observation effect' describes the alteration of dietary behaviours during the period of study (Stubbs et al., 2014). The prevalence of low energy reporting has been illustrated in the NHANES study (n=~63,000), in which estimates of EI for 67.3% of women and 58.7% of men were considered to be physiologically implausible (i.e. EI <1.35 x RMR), leading to the conclusion that these measures are 'pseudo-quantitative' (Archer et al., 2013), though this plausibility cut-off, which is discussed below, is highly arbitrary (Stubbs et al., 2014). In a pooled analysis of 5 large validation studies conducted on variable U.S. samples, less than 10% of the variance in EI (assumed to equal TDEE from DLW) was explained by the values obtained by self-report (Freedman et al., 2014).

Several cut-offs have been proposed to estimate the plausibility of reported EI. The 'Goldberg cut-off 1' considers a value of 1.35 x RMR as the minimal plausible EI (Goldberg et al., 1991). This is a widely used approach but is subject to criticism (Stubbs et al., 2014) owing to its arbitrary nature and

limited capability to identify misreporting. A subsequent, more involved cut-off (cut-off 2) was proposed which considers the number of days over which self-report is obtained, coefficients of variation for EI, physical activity level and sample size and was designed to detect both high and low energy reporting (Goldberg et al., 1991). Whilst these methods were designed to assess the plausibility of self-report data, and may identify the most substantial misreporting, their application to identify misreporting assumes that each subject is either misreporting or not and therefore fail to identify those misreporting to a lesser degree or misreporting particular macronutrients.

Theoretically, if strong and consistent predictive characteristics associated with the degree of misreporting could be identified, it may be possible to correct the biases associated with self-report measures. Both gender (Murakami & Livingstone, 2015) and body mass index (BMI) (Rasmussen et al., 2007) have been associated with the magnitude of misreporting of EI. Some studies have reported that females misreport to a greater degree than males (Mendez et al., 2004; Pfrimer et al., 2015). However, others find no differences between genders (Rasmussen et al., 2007). Regarding BMI, those with a higher BMI have tended to under-report to a greater extent (Mendez et al., 2004; Rasmussen et al., 2007), although some studies also report no relationship between BMI and magnitude of misreporting (Asbeck et al., 2002). The identification or prediction of misreporting on the individual level is currently impossible as reliable predictors remain elusive and therefore self-report EI (not necessarily diet composition) is of limited use to quantitative energy balance research.

### 1.2.2 Quantifying energy intake with intake balance methods

Thermodynamic principles applied to human physiology make it possible to solve the energy balance equation if two of the three components are known. The two measured components do not matter necessarily; it is possible to vary EI until $\Delta ES$ = 0, thereby establishing maintenance requirements (Heymsfield et al., 2017). This approach is practically challenging and only provides information on the subject's maintenance requirements. A more widely used approach is to estimate the $\Delta ES$ and EE and to solve for EI. This technique, commonly referred to as the 'intake-balance method' requires two body composition measures, from which the energy density of the $\Delta ES$ can be estimated and if EE of the subject is known, EI can then be calculated.

Implementing the intake-balance method using DLW and Dual-energy x-ray absorptiometry (DEXA) can be considered a gold-standard for deriving EI (Racette et al., 2012). However, the error associated with the intake-balance method is not 0, and analytical and physiological errors are often assumed to sum to ~5% for DLW (Black & Cole, 2000; Trabulsi et al., 2003) and ~1% for DEXA (Sanghvi et al., 2015). The analysis for these techniques can also take several months depending on the facility, making real-time analysis impossible. With this in mind, research groups are exploring alternative methods for the quantification of the components of the energy balance equation. Activity monitors are recognised as the most viable opportunity to accurately estimate EE in large groups of free-living individuals. Shook and colleagues investigated whether the expensive DLW measures can be replaced with TDEE estimates from a SenseWear armband mini (SWA) which is a research-grade, arm-worn activity monitor (Shook et al., 2018). The results were promising, with TDEE, PAEE and therefore EI differing by <2 kcal/day on average, though an $R^2$ value of .44 implies imperfect agreement (Shook et al., 2018). The observed accuracy in EI estimates is almost entirely dependent on the SWA, given the smaller error associated with the DEXA method. It is therefore important to note that the errors increased with increasing TDEE, with a bias of 160 kcal/day relative to DLW in the highest TDEE group, so this method will require further validation in independent samples. Critically, the SWA is no longer in production (Gibbs & Davis, 2018) and is therefore unlikely to offer a scalable solution for future studies. However, the work of Shook and colleagues provides evidence that wearables could offer an alternative to DLW for energy balance studies.

A mathematical strategy to estimate EE during a positive energy balance has been proposed and this method could offer a cost-effective alternative for DLW measures in the typical intake-balance protocol. This model, proposed by Gilmore et al., accounts for increases in DIT and tissue deposition and has been used to estimate the EI of subjects in a metabolic ward study (n=8) and an outpatient free-living DLW validation (n = 35). No differences were observed in the inpatient study but a small significant underestimation was seen in the outpatient study (Gilmore et al., 2014). This approach has been criticised for failing to account for all the necessary components of EE, such as the increased EE attributable to physical activity during overfeeding, increased maintenance energy requirements and NEAT, all of which likely contribute in different degrees to EE during weight gain (Hall, 2014).

**Table 1.1** Methodologies to estimate energy or food intake.

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| **Food records** | A log of all foods eaten is recorded by a subject. Food records require coding by trained personnel. | Up to 1 week | • Not subject to recall bias<br>• 7-day food record provides a quantitative estimate of energy and macronutrient intake<br>• May capture diet variability | • Expensive<br>• Requires trained dietitians to code responses<br>• Subject to experimenter error<br>• Large participant burden<br>• Subject to misreporting |
| **Food recalls** | Food and drink consumption over a predefined period (i.e. previous day) is determined by interview or software. | 1 day | • Ease of administration facilitates use in large studies<br>• Can be conducted by telephone or by computer programmes | • Subject to recall bias<br>• Subject to misreporting<br>• Requires trained dietitians or software to code responses<br>• Subject to experimenter error<br>• Repeated measures required to assess variability |
| **Food frequency questionnaire** | A list of foods is presented and subjects report the frequency of consumption during a specified period (up to one year). | Up to 1 year | • Ease of administration facilitates use in large studies<br>• Standardised questionnaires can be scanned electronically<br>• Cheap<br>• Low participant burden | • Subject to recall bias<br>• Subject to misreporting<br>• Does not provide quantitative estimates of energy intake |
| **Photography methods** | Respondents photograph meals before and after eating/ Images may be accompanied by a marker in the images (for | Variable | • Objective estimation of portion size<br>• Allows for the identification of data entry | • Privacy considerations<br>• Requires cameras or smartphone<br>• Requires trained dietitians or software to code responses |

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| | portion size estimation). | | errors<br>• Not subject to recall bias | • Subjects may forget to photograph meals |
| **Observed intake** | Trained researchers observe food intake in a controlled clinical environment, often food is available ad libitum | Variable | • High level of objectivity and accuracy | • Expensive<br>• Low ecological validity<br>• Subject to Hawthorne effect |
| **Mathematical models** | Linearised mathematical models of body weight dynamics are used to estimate the change in energy intake from baseline maintenance requirements from repeated body weight measures | Years | • Cheap<br>• Very low participant burden<br>• Not subject to recall bias<br>• Not subject to misreporting | • Complicated to implement<br>• Associated with numerous assumptions which may not hold for all individuals<br>• Validation studies are rare |
| **Biomarker or metabolomic methods** | Samples are collected (e.g. saliva, urine) that are associated with dietary intake or the status of these nutrients | Variable | • Highly accurate for some dietary components (i.e. nitrogen intake)<br>• Not subject to recall bias<br>• Not subject to misreporting | • Can not provide a quantitative estimate of total energy or dietary intake<br>• Technical expertise required<br>• Large participant burden<br>• Expensive<br>• Participant and lifestyle factors influence metabolites<br>• Validity of metabolomic approaches is unclear |

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| **Intake-balance method** | Energy intake is calculated based on the change in energy stored (fat mass and fat-free mass) and average daily energy expenditure. Body composition is assessed at two points and energy expenditure is assessed in the interim. | 14 days | • Highly precise and accurate (DLW and DEXA)<br>• The method can be applied with wearable devices<br>• Not subject to recall bias<br>• Not subject to misreporting | • DLW is expensive<br>• Provides no information on the nutrient intake<br>• Provides average energy intake over the measurement period<br>• High participant burden |

## 1.3 Measure of energy expenditure

### 1.3.1 Quantifying energy expenditure with self-report measures

Similar to the measurement of EI, a variety of self-report tools for EE and physical activity are available. Methods involve variants on 7-day and yearly recall questionnaires, as well as an array of logs and diaries (Helmerhorst et al., 2012). Many of the questionnaires involve conversion steps between the subject's qualitative reports and the energetic work of activities and this is achieved by looking up values in the compendium of physical activities (Ainsworth et al., 2011) which are population estimates with limited applicability at the individual level (Hills et al., 2014). A 2012 review investigating the validity of 96 existing physical activity questionnaires relative to objective measures reported maximal correlation coefficients of 0.76 for reliability and 0.41 for validity (Helmerhorst et al., 2012). A recent comparison of self-report measures of physical activity converted to estimates of EE (n=78) with DLW showed a significant difference for PAEE of 414.6 kJ/day (range: 78.7, 750.5 kJ/day), though in this case, the mean may be deceptive, as none of the tested measures were correlated with DLW across a population or individually (Sharifzadeh et al., 2020).

### 1.3.2 Quantifying energy expenditure with wearables

The growth in wearable technologies and algorithmic approaches in recent years is revolutionising EE assessments, with potentially enormous implications for biomedical sciences (Wright et al., 2017). Using wearable devices to monitor human behaviour has been a research interest for many years (Ceesay et al., 1989) but very recent technological and engineering advancements have allowed current wearables to incorporate GPS sensors, respiratory sensors, heat sensors, goniometers, accelerometers, heart rate monitors and gyroscopes (Yang & Hsu, 2010). All of which can generate extremely detailed time-series datasets of human movement and physiological signals. The following section details recent advances in accelerometer-based or physiological devices.

#### 1.3.2.1 Metabolic equivalents

In most predictive accelerometry research the aim is to predict metabolic equivalents (METs) rather than an absolute value of EE. Much of EE is determined by RMR, which is determined by the subject's body composition, age, gender, and environmental conditions (Hopkins & Blundell, 2016) rather than mechanical work. One MET is the $VO_2$ at rest, which has generally

been assumed to be 3.5 mL/$O_2$/min/kg body weight (Hills et al., 2014). Standardisation of physical activities in this manner allows consistency and comparability of the energy cost for subjects of different weights and the compendium of physical activities provides a comprehensive list of METs estimations for numerous physical activities (Ainsworth et al., 2011) which can subsequently be used to derive EE. Despite the prevalence of the 3.5 mL/$O_2$/min/kg assumption, research is indicating that this may be a poor reflection of the true resting EE (Byrne et al., 2005) as wide variation in body composition and size exists. This assumption may be invalid, particularly in people with obesity (Hills et al., 2014).

#### 1.3.2.2 Heart rate methods

The monitoring of heart rate has been used for the estimation of EE because of the relatively linear relationship it shows with $VO_2$ above moderate-intensity activity, but this relationship is not observed at lower intensities (Strath et al., 2000). Tracking heart rate is not costly and is highly portable, facilitating the assessment of free-living individuals (Leonard, 2003). Unfortunately, due to variability in cardiorespiratory fitness, age and genetics, the linear relationship (both the intercept and slope) between $VO_2$ and heart rate varies significantly between individuals (Leonard, 2003). The potential differences in heart rate at various levels of physical activity is so large that individual-level calibration procedures appear to be necessary (Brage et al., 2007).

Considering this, the pioneering ‘flex-HR’ method was developed (Spurr et al., 1988). Determination of the flex-HR parameters requires an individual calibration protocol in which $VO_2$ and heart rate are measured simultaneously. This protocol defines the ‘flex point’, the slope and intercept of the heart rate and $VO_2$ regression for a single subject (Welk, 2002). The flex point is defined as the average of the highest resting heart rate value and the lowest exercising heart rate (Leonard, 2003) which serves as a cut-point above which $VO_2$ and EE can be inferred from the regression parameters. Below the flex point, average resting $VO_2$ is often defined based on the average of resting postures, also measured during the calibration protocol (Spurr, 1990). With these parameters, a $VO_2$ prediction is obtainable given a measured heart rate. Early validations of the method against whole-body calorimetry showed reasonable agreement across a range of activities ($R^2$ = 0.87) and an inter-individual variation between 20% and -15% for TDEE (Spurr et al., 1988). These results were subsequently

replicated (Ceesay et al., 1989) and have been validated against DLW in adults (Livingstone et al., 1990).

A central limitation of the method is the time-consuming nature of the calibration protocols, which must be repeated for every subject (Brage et al., 2007). Attempts have been made to estimate the flex point, the intercept and slope of the regression line based on more easily obtained values such as sex, age, body weight, BMI, FM% and sitting heart rate (Rennie et al., 2001). This study reported that physical activity level estimated with these parameters was highly correlated (r = 0.82) with the estimates obtained using the calibrated values in a sample of 97 adults, indicating a potential means of bypassing the burdensome calibration process.

Further limitations of the flex-HR method include the lack of correlation between heart rate and EE at low-intensity activity and because of this EE is assumed for sedentary behaviours, which represents most of the day for western adults. The method is also subject to variability in heart rate response to environmental conditions, both of which are likely to reduce the accuracy and precision of the method (Welk, 2002). Another limitation is the variation seen in the relationship between heart rate and $VO_2$ depending on musculature recruited (Hills et al., 2014). As technology has developed the flex-HR method has become less frequently used in research, however, the underlying principle of linearity plays an important role in more modern EE estimation methods (Brage et al., 2015).

#### 1.3.2.3 Accelerometers

Micro-electromechanical system technology underlies most modern accelerometers and this has facilitated the development of tiny yet accurate accelerometers capable of measuring both static and dynamic accelerations (Plasqui, 2017). Accelerometers represent the most common sensor within wearable devices, with the vast majority incorporating an accelerometer in 2016 (de Arriba-Pérez et al., 2016) and many of these modern devices register movement in three planes (anteroposterior, mediolateral, and vertical) (Chen & Bassett, 2005).

As there is a nearly linear relationship between the energy cost of muscular force generation and acceleration, the measurement of acceleration can be used to infer the intensity of activity (Ridgers & Fairclough, 2011). Acceleration is typically measured in gravitational acceleration units (g) where 1 g is equal to 9.8 meters/second$^2$. The raw g output is commonly converted to ‘counts’ and is expressed relative to a unit of time, ‘epochs’

(Chen & Bassett, 2005). From here, counts per epoch are converted to more relevant parameters, such as steps, activity classification or EE (Hills et al., 2014). Before the derivation of activity metrics, the raw acceleration signal passes through several steps including filtering to minimise artefacts in the signal, an integration step, an extraction step over defined time frames and the application of algorithms or cut-points to derive the variable of interest (Chen & Bassett, 2005). The assumptions and methods used at each of these steps vary substantially between manufacturers and analysis software, which makes comparability between studies challenging (Plasqui, 2017).

Deriving estimates of PAEE from movement signals is of great research interest and the aggregated signal from tri-axial accelerometers has a relatively linear relationship with EE in many activity modalities (Crouter, Churilla, et al., 2006), making it a potentially useful tool in the assessment of EE. A simple linear model was proposed over 20 years ago which is used to transform counts to EE (Freedson et al., 1998). The proposed model, which was derived from data collected during a treadmill protocol, generalises well to ambulatory activity but not to non-ambulatory activity such as household tasks (Hendelman et al., 2000) or running at very high velocities (Kozey et al., 2010). Prediction equations were subsequently developed on more diverse training data; Swartz et al. (2000) used a protocol comprised of walking tasks and a number of lifestyle activities to develop regression equations (Swartz et al., 2000), which explained 34% of the variance in a hip and wrist combined model. Two-regression models were subsequently developed which utilise two different models, depending on the measured counts. Crouter et al used cut-offs (>/<10 coefficient of variation of counts/10s) to select either a walk/run model or a lifestyle model, which led to considerable improvements in estimations of METs, compared with a single model (Crouter et al., 2010). Crouter's refined method was observed to have no significant difference from measured METs for any activity except cycling when compared to an indirect calorimeter (Crouter et al., 2010).

A comprehensive comparison amongst these algorithms and 9 other equations was conducted in a protocol consisting of treadmill activities and daily living tasks; overall these models were shown to be inadequate for EE prediction as well as classification across a range of intensities (i.e. light, moderate or vigorous) (Lyden et al., 2011). To advance this area of research it is important to consider the sources of error and inconsistency. Firstly, EE is estimated based on time spent at various intensities and there is no standardisation of the number of 'counts' required to define each intensity

(Hills et al., 2014). Many cut-points have been developed for different populations and have limited applicability to others (Troiano et al., 2014). The incorporation of physiological signals with accelerometer data offers an exciting opportunity to improve the modelling of EE.

#### 1.3.2.4 Multisensory approaches

Both heart rate and accelerometer methods have method-specific limitations (Schutz et al., 2001) and combining numerous sensor outputs could improve estimates of free-living EE. Combination approaches can refer to either multiple accelerometer sensors or accelerometers combined with physiological sensors and both of these approaches are considered below.

The intelligent device for energy expenditure and activity (IDEEA) has been used to derive EE and activity through data obtained through multiple accelerometers attached to multiple body sites (sternum, thighs, and both feet) (Zhang et al., 2004). In one validation study, the IDEEA monitor (using measured RMR) demonstrated a significant underestimation relative to a whole room calorimeter (~0.9 MJ) (Whybrow, Ritz, Horgan, & Stubbs, 2013). Although the device has been applied to estimate EE in free-living participants in the DiOGenes study (Larsen et al., 2010), the cumbersome nature of attaching multiple sensors has probably contributed to the limited usage of the IDEEA and this must be an important consideration for research utilising wearables.

The linear relationship between heart rate and $VO_2$ is not observed at low intensity, and accelerometers generally have little ability to distinguish the higher energy cost associated with factors such as carrying a load or walking at an incline. In theory, this means that they can complement each other (Hills et al., 2014). The Actiheart (Brage et al., 2005) is a chest-worn device which is capable of measuring acceleration and heart rate and was novel in that it combines heart rate sensing and acceleration in one device (Crouter et al., 2008). The Actiheart's algorithm has been published and predictions are generated via a branched model; based on the observed acceleration and heart rate, EE is estimated by different linear models utilising accelerometry only, heart rate only or a combination of both (Brage et al., 2004). The parameters of the prediction model can be derived for each subject in a submaximal test and if this is not possible, group-level estimates can be used. However, research has generally shown that the accuracy and precision of the Actiheart are improved when the device is individually calibrated (Brage et al., 2015). A free-living validation of the Actiheart in a Cameroonian population of 33 adults demonstrated a small mean bias of

−5.4 and −9.1 kJ/kg/day relative to DLW for the individual and group level calibrations, respectively (Assah et al., 2011). Another DLW study reported correlations of r = 0.67 and non-significant bias with individually calibrated models in 46 healthy adults (Brage et al., 2015).

The SWA (BodyMedia, Inc., Pittsburg, PA) has been available in several forms for many years (Fruin & Rankin, 2004). The device is worn on the upper arm and includes a triaxial accelerometer (newer models only), a galvanic skin response sensor, skin and air temperature sensors (Slinde et al., 2013), all of which are combined in a proprietary algorithm to derive EE predictions at the minute level (Santos-Lozano et al., 2017). The proprietary algorithms (which have been incrementally updated) and alterations to hardware configurations are known to produce different estimates of EE (Bhammar et al., 2016). Under free-living conditions, the device has generally agreed well with DLW with correlations of r = 0.66-0.80 reported (Farooqi et al., 2013; Johannsen et al., 2010; Koehler et al., 2011) and near-perfect agreement at the group level in one study (Shook et al., 2018). Concerning accuracy in specific activities, a substantial bias of > 1.5 METs has been observed in high-intensity ambulatory exercise (Santos-Lozano et al., 2017) and correlations with indirect calorimetry varied from r=0.39 to 0.93 in a running protocol (Drenowatz & Eisenmann, 2011). The SWA has been extensively used in physical activity research and the large body of validation literature suggests a good accuracy at the daily level in moderately active adults (Myers et al., 2019; Nymo et al., 2018; Shook et al., 2018). The tool is minimally invasive, which undoubtedly contributes to its popularity relative to more cumbersome devices such as the Actiheart or IDEEA. Unfortunately, these devices and others are limited by the recording capacity. They must be returned to the laboratory for charging or download, which limits their use over long durations. Many of these devices can collect data over periods of 14-days maximally. Thus, researchers effectively assume that the period of observation is representative of the behaviours in the weeks or months where the device is not worn, which is likely to vary with social and environmental factors (Aspvik et al., 2018; Chan et al., 2006; Merchant et al., 2007; Shiroma et al., 2019).

#### 1.3.2.5 Commercial activity monitors

A limitation of research-grade devices is both scalability and limited duration of the measurement. By contrast, commercial devices are cloud-connected through Bluetooth and smartphone apps, allowing a continuous upload from a device, which provides a constant stream of data into an application

programming interface (API). Furthermore, commercial monitors are designed to be minimally burdensome and many are now worn on the wrist, which is likely to further promote compliance (Troiano et al., 2014).

Sales and production of these devices have been increasing drastically despite concerns regarding their accuracy. Wallen and colleagues, in a simple laboratory protocol involving resting, ambulatory and cycling tasks reported a substantial error in EE estimates relative to indirect calorimetry (9–43%) for several wearables including the Apple Watch and Fitbit Charge HR (Wallen et al., 2016). An earlier systematic review in this area concluded that commonly used activity monitors typically show low validity but excellent reliability (Evenson et al., 2015). A later review focussing exclusively on Fitbit devices found that across 18 studies, average errors within 3% were rarely achieved and generally an overestimation of EE during activity behaviour was observed (Feehan et al., 2018). Variation in ambulation (speed, incline, surface types) impacted the accuracy of the devices and most resting comparisons showed that Fitbit devices underestimate EE (Feehan et al., 2018). Few DLW comparisons have been made for Fitbit devices, but one suggests an underestimate of ~ 7% in TDEE (Murakami et al., 2016)

It appears Fitbit and other commercial monitors are limited in their ability to predict EE, however, a redeeming characteristic of these devices appears to be the accuracy achieved in estimating heart rate. One study reported that the majority of estimates falling within 5% of a gold-standard measurement for a range of activities (Shcherbina et al., 2017) and another reported modest errors of ~5 bpm in a cycling based protocol (Benedetto et al., 2018). However, differences in BMI and skin tone may influence measurement error (Stahl et al., 2016). A characteristic of commercial monitors is that the rate of hardware and software updates is regular compared to research-grade devices and silent changes in algorithms are therefore possible.

#### 1.3.2.6 Statistical combination approaches

The limitations of simple linear models applied to count data for estimation of EE are well documented. In recent years, fuelled by computational advancement, data availability and the implementation of complex algorithms in high-level programming libraries, machine learning approaches have emerged as a promising opportunity for the estimation of EE. In studies applying these methods, signals are extracted, and algorithms are used to learn complex, nonlinear functions mapping sensor data to EE or activity category (Sardinha & Júdice, 2017). These techniques can enhance the

precision of signal processing and movement assessment (Farrahi et al., 2019).

In one of the first studies in the area, researchers trained artificial neural networks to predict METs and reported a root-mean-squared error (RMSE) of ~ 1.2 METs relative to an indirect calorimeter, which reduced the EE estimation error substantially relative to the Crouter two-regression model (Staudenmayer et al., 2009). Ellis and colleagues compared estimates derived from random forests trained on data from hip or wrist devices and were one of the first to incorporate heart rate in their models. The use of heart rate improved MET estimation, achieving a RMSE of ~1 MET relative to a portable indirect calorimeter in a series of household, resting and exercise tasks (Ellis et al., 2014). Similar findings have been reported with neural networks applied to accelerometer data collected from the hip, thigh and wrist in a semi-structured protocol; the authors report the smallest error at the thigh attachment site (RMSE = 1.04 METs), which was closely followed by wrist and hip worn devices, in terms of RMSE (Montoye et al., 2015). The authors compared linear and non-linear (artificial neural networks) models for predictive accuracy in a follow-up study. The performance advantage of the complex approaches was only seen when processing wrist accelerometer data, correlations relative to indirect calorimetry were up to r = 0.84 for the neural networks and r=0.73 for linear models (Montoye, Begum, et al., 2017). Indeed, the wrist experiences inconsistent and variable accelerations when compared to other tested body sites and this likely requires more complex modelling approaches.

Research has also treated estimating EE and physical activity as an intensity classification problem by predicting an activity class (sedentary, light, moderate and vigorous), given some sensor input. In the most comprehensive study to date, Farrahi et al collated 5 independent studies and trained a series of artificial neural networks to classify sedentary, light or moderate-to-vigorous physical activities based on the accelerometer signals. High predictive accuracy was observed when the datasets were combined, reaching up to 90.7% accuracy but substantial performance degradation was observed when these models were applied to independent datasets (Farrahi et al., 2020). Whilst time in activity categories is an important metric on which interpretable public health guidelines are based (Ostendorf et al., 2018), it does not facilitate estimates of TDEE. However, segmentation of activities into specific categories has been a recurring theme in

accelerometer research and it may offer an opportunity to refine TDEE estimates.

Machine learning approaches to date have relied on small training datasets, the neural network architectures have had relatively few parameters and attempts at tuning hyperparameters have not been particularly comprehensive, all of which limit the capacity for feature extraction and predictive accuracy (Cao et al., 2018). Another important observation of these models to date is a tendency for predictions to regress towards the mean of the training data and therefore result in an overestimation of the energetic cost of resting/sedentary epochs (O'Driscoll, Turicchi, Hopkins, et al., 2020). Overall, these methods are in their infancy within computational bioenergetic and energy balance modelling fields and despite promising laboratory results, their free-living potential is not yet clear.

### 1.3.3 Gold-standard measure of energy expenditure

#### 1.3.3.1 Direct and indirect calorimeter methods

Direct calorimetry, indirect calorimetry and non-calorimetric methods represent the categories of methods used to estimate EE (Westerterp, 2015). Neither indirect nor direct calorimetry can be considered suitable for continuous, longitudinal free-living measures, and the intricacies of these methods are beyond the scope of this chapter. However, they serve as important validation and development tools for many of the methods discussed later and are therefore introduced below.

Historically, the quantification of EE in humans has targeted the measurement of heat production, as energy utilised for metabolic purposes is ultimately lost as heat (Lam & Ravussin, 2017). Direct calorimeters capture this heat production within a whole-room or body-suit calorimeter in which adults can reside in for short metabolic studies (Shephard, 2017). Direct calorimeters historically have offered a highly accurate means of quantifying TDEE but the extremely high cost to build and maintain mean that very few facilities exist worldwide (Tamura, 2019). Indirect calorimetry is the method most commonly used for the quantification of EE and does not directly measure heat production, but instead the oxidation rates of dietary substrates through the measurement of consumption of oxygen ($VO_2$) and/or production of carbon dioxide ($VCO_2$) (Haugen et al., 2007). The $VO_2$ and $VCO_2$ reflect macronutrient oxidation rates and therefore can be used to infer heat production (da Rocha et al., 2006). When $VCO_2$ and $VO_2$ are measured

EE can be calculated through the application of the Weir equation (Weir, 1949) or alternatives (Tamura, 2019).

Laboratory-based indirect calorimetry methods, in which the participant wears a face mask connected to a metabolic system, are widely used oweing to a high level of precision and accuracy but are associated with a limited duration of measurement due to the discomfort of the apparatus (Westerterp, 2015). They are extremely accurate and precise over epochs as short as 15 seconds (Tamura, 2019), making them ideal for validation of wearable activity monitors (Chowdhury et al., 2017) and the development and validation of new prediction algorithms for EE (Staudenmayer et al., 2009). It is also possible to apply the theory of indirect calorimetry to portable devices in which such systems can be used with backpacks and facemasks to facilitate EE measurement outside of laboratory environments (Gupta et al., 2017). However, the duration of the measurement is still restricted to a few hours (Lam & Ravussin, 2016) and the precision and accuracy are poor when compared to more established stationary systems (Tamura, 2019). Refinement of these technologies will likely facilitate more ecologically valid data for algorithmic development.

#### 1.3.3.2 Doubly labelled water

Estimation of TDEE with the DLW method is a form of indirect calorimetry, in which TDEE is derived from estimates of $CO_2$ production, rather than the measurement of heat production (Lanningham-Foster et al., 2005). The method was first applied to humans in 1982, approximately 30 years after its invention, mostly attributable to the high cost of the method (Speakman, 1998) and it has been widely applied in human energetics research since. The DLW method is currently considered the gold-standard for the assessment of free-living TDEE and in combination with gold-standard measures of body composition (see **section 1.4.3**), can give a gold-standard EI estimate in free-living subjects (Dhurandhar et al., 2015). The method has facilitated investigations into the rates of TDEE in elite cyclists participating in the Tour de France (Westerterp et al., 1986), the energetic cost of Antarctic expeditions (Stroud et al., 1997) and Hadza hunter-gatherers living in northern Tanzania (Pontzer et al., 2012).

The DLW method is based on the premise that two labelled isotopes, deuterium ($^{2}H$) and oxygen-18 ($^{18}O$), equilibrate with total body water. Both of the isotopes are eliminated over time but because $^{2}H$ exits the body as water exclusively and $^{18}O$ equilibrates with the body water and the carbon dioxide pool through the carbonic anhydrase reaction, it leaves the body as

both $H_2O$ and $CO_2$. The difference in the elimination rate of the two isotopes is an estimate of $CO_2$ production (Westerterp, 2017). The method requires the collection of a baseline urine sample and subsequently a bodyweight specific dose of DLW is administered, marking the start of an assessment (Park et al., 2014). The duration of the measurement period is dependent on the water turnover, with higher water turnovers (i.e. endurance athletes) having a reduced period of measurement for a given isotopic dose (Ekelund et al., 2002; Schoeller et al., 1986). During the free-living observation period, participants provide urine or bodily fluid samples and the rates of elimination of the isotopes are calculated for the subjects by mass spectrometry (Delany, 2012). Once the $CO_2$ production is known, EE can be calculated through the Weir equation (Weir, 1949) and divided by the number of days to give the average TDEE (Livingstone et al., 2003). Further methodological discussion, including assumptions of the DLW method, is provided in **section 3.4.2.4**.

Early validations relative to a respiratory chamber reported that the method has a precision between 2-8% (Schoeller et al., 1986) and subsequent evaluation in lean and obese subjects reported an average difference of <3%, with greater underestimates observed in the most overweight subjecs (Ravussin et al., 1991). More recently, De Jonge, et al. demonstrated the accuracy of the measure during caloric restriction, with a mean difference of 1.3 ± 8.9%, relative to a respiratory chamber in subjects following 3 weeks of caloric restriction (de Jonge et al., 2007). The CALERIE study has provided a novel opportunity for investigation of the method; Wong et al. showed turnover rates were repeatable to within 1% and 5% for $^{2}H$ and $^{18}O$, respectively, providing strong evidence that the method is suitable for longitudinal energy balance research (Wong et al., 2014).

The excellent precision and accuracy stated in these studies are dependent on meeting the methodological and theoretical assumptions of the method. For example, the $CO_2$ production rate and the size of the bodily water pool are assumed to be constant for the duration of measurement (Speakman, 2018). Furthermore, in many experimental conditions, the respiratory quotient of the diet is assumed rather than measured. The respiratory quotient value is variable depending on the dietary patterns, physical activity level and energy balance of the subject (Elia, 1991) and recent work suggests that failure to account for this in low carbohydrate diets may bias DLW estimates (Hall et al., 2019). Inconsistency in methodologies (i.e. multiple urine samples vs two-point urine samples) may also limit

comparability between findings of independent studies. Relative to two-point samples, daily urine sampling has been reported to improve precision and accuracy (Berman et al., 2020).

Some further limitations of the DLW method must be acknowledged. First, it is currently impossible to validate field assessments and validation studies must be conducted in laboratory settings. Second, the substantial costs of isotopes and analysis limit the number of participants it is feasible to study. Third, the relatively short period of assessment (7-21 days) is often assumed to represent habitual behaviours outside the measurement period. It is impossible to eliminate the Hawthorne effect, whereby the act of observation leads to behavioural changes in subjects (McCambridge et al., 2014) and therefore researchers must exercise caution in generalising beyond the observation period. Perhaps the most significant limitation of the DLW method is that it only provides an average TDEE value. Physical activity is highly variable in terms of intensity, duration and frequency throughout the day. Different patterns of physical activities can produce different metabolic (Gonzalez et al., 2013) and appetitive/EI (Höchsmann et al., 2020) responses, which the DLW method cannot be used to study. Many other methods of EE assessment are not associated with the same limitations but carry their own limitations. A summary of commonly used methods is shown in table 1.2.

**Table 1.2** Methodologies to estimate EE or physical activity.

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| **Self-report activity logs** | A log of all activity is kept by subjects. Logs are coded by researchers. | Weeks – months | • Fast<br>• Cheap<br>• Low participant burden | • Subject to misreporting |
| **Direct observation** | Participants are observed in a laboratory by researchers or they wear a camera, which is converted to activities by researchers. | Hours - Days | • Highly accurate estimates of time in physical activity<br>• Biomechanical software can assist in coding activities | • High researcher burden<br>• Verification required<br>• Hawthorne effect may influence behaviour<br>• Limited validation studies<br>• Translating activities to energy expenditure may be inaccurate |
| **Heart rate monitoring** | Participants perform a calibration procedure in which heart rate and $VO_2$ are measured continuously. A linear model is used to estimate energy expenditure based on measured heart rate. | 10-14 days, determined by battery life | • Not subject to misreporting<br>• High ecological validity | • Requires individual calibration<br>• Calibration parameters may depend on the activity performed<br>• High participant burden<br>• Sensors may produce erroneous data<br>• High error at the individual level |
| **Research-grade accelerometers** | Activity monitors are initialised by researchers and sent or given to participants. Activity monitors track movement in 1-3 axes and outputs | 10-14 days, determined by battery life | • High ecological validity<br>• Raw data may be available for modelling | • Proprietary algorithms in some devices limit understanding of outputs<br>• Most devices estimate energy |

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| | are converted to physical activity or energy expenditure by cut-point methods. Some devices also incorporate physiological sensors. Activity monitors are worn on the wrist, hip, chest or back. | | | expenditure on accelerometer signal only |
| **Commercial-grade wearables** | Activity monitors track movement in 1-3 axes and typically heart rate. Outputs are converted to physical activity or energy expenditure by proprietary algorithms. Activity monitors are most commonly worn on the wrist. | Years | • Cloud connectivity facilitates long term measurements<br>• Low researcher burden<br>• Low participant burden<br>• High ecological validity | • Proprietary algorithms in some devices limit understanding of outputs<br>• Hardware and firmware may be updated regularly<br>• Manufacturer estimates are poor in many cases<br>• Motivational aspects of devices may alter behaviour<br>• Missing data is likely over long periods of measurement |
| **Direct calorimetry** | Whilst a subject is enclosed in a chamber, all heat transfer including radiation, convection, conduction, evaporation is measured. Heat production is typically obtained by observing the differences in air temperature and humidity between the input and output. | Up to 1 Week | • Directly measures heat production<br>• High accuracy and precision | • Accuracy may decrease for high-intensity activity<br>• Technical expertise required<br>• High participant burden<br>• Low ecological validity |
| **Whole-room respiratory chamber** | Subject resides in a chamber and air of known composition is pumped in. The outflowing air is analysed to determine the oxygen consumption and carbon dioxide | Up to 1 Week | • High accuracy and precision | • High cost to maintain and install<br>• High participant burden<br>• Low ecological validity |

| Method | Description | Observation period | Advantages | Limitations |
|---|---|---|---|---|
| | production. | | | |
| **Metabolic carts** | Oxygen consumption and carbon dioxide production are measured via a facemask or ventilated hood, which is analysed by the metabolic system, | Hours | • Moderate to low maintenance costs<br>• High accuracy and precision<br>• Low response time (<30s)<br>• Semi-portable | • High purchase cost<br>• Not suitable for long-term measures<br>• Low ecological validity |
| **Doubly labelled water** | Subjects are dosed with stable isotopes ($^{2}H$ and $^{18}O$), these isotopes equilibrate with hydrogen and oxygen in total body water. Urine samples are collected and analysed to determine the differential elimination rates of the isotopes, which indicates carbon dioxide production, and energy expenditure. | Up to 3 weeks, depending on the water turnover | • The gold standard for energy expenditure measures in free-living environments<br>• Applied in a wide range of subjects and environments<br>• High ecological validity | • Provides average energy expenditure over the observation period<br>• Accuracy and precision depend on a series of assumptions and calculations<br>• High participant burden<br>• High cost for isotopes<br>• Technical expertise required for analysis |

## 1.4 Measure of energy storage

Since energy metabolism in humans adheres to the principle of energy conservation, a change in the composition of the body has an energetic value, which is equivalent to the energy imbalance. Thus, a change in the weight or composition of the body is used to infer the energy balance status of individuals.

### 1.4.1 Bodyweight

Approximately constant body weight over time (i.e. ± 1 kg) implies EI ≈ EE and weight stability is therefore used to illustrate energy balance (Heymsfield et al., 2017). However, this provides no detail on the degree of rate of EE and EI (Hand & Blair, 2014) or the energetic cost of weight change, which is of interest to most weight management studies. Arguably the most prevalent method for approximating the energy density of weight change in clinical settings has been the '3500 kcal rule', which erroneously assumes an energy deficit or surplus of 3500 kcal will lead to 1 pound of weight change and is derived from historic work approximating the energy content of fat (Hall, 2008; Wishnofsky, 1958). The uniformity of this rule predicts continued weight loss without a plateau (Thomas et al., 2013) and ignores factors influencing the composition of weight change, including the initial body composition, rate of weight change (Heymsfield et al., 2014) or the short-term fluctuations in glycogen and water.

A frequently used coefficient for the energetic cost of weight change is 7.4 kcal/g (Gilmore et al., 2014; Tataranni et al., 2003), which considers the partitioning of energy imbalance between FM and FFM. Racette et al., working with a subsample of 40 subjects in the CALERIE study, investigated whether body weight can be used to approximate $\Delta ES$. The authors regressed body weight over time and used a coefficient of 7.4 kcal/g to approximate the energy cost of weight change (Racette et al., 2012). This approach was not significantly different from the reference Δbody composition by DEXA for 4-week changes; indicating the potential for weight alone to provide reasonable estimates of $\Delta ES$. This coefficient is likely to vary between individuals, direction of weight change and probably becomes less valid at the extremes of energy imbalance, body composition and exercise. For example, a value of 8.4 kcal/g was calculated in an adult overfeeding study (Gilmore et al., 2014). Though constant coefficients such as these fail to recognise the distinction between the different phases of weight change. The phases of weight loss and their physiological

characteristics have been comprehensively examined previously (Heymsfield et al., 2011). In summary, the initial phase of weight change is characterised by a rapid rate of change in FFM (via water, electrolytes and nitrogen losses) and as weight loss progresses towards a second phase, the provision of energy is increasingly met by stored lipid (Heymsfield et al., 2011), therefore the energy content associated with a weight change is variable.

### 1.4.2 Mathematical models

Many thermodynamic models have been proposed which can be applied to quantify body weight dynamics in humans and these models can be solved for EI (Thomas et al., 2019). The advantage of these approaches is that they do not require laboratory measurements and can be easily applied to large groups of participants with a minimal associated cost.

Thomas et al. proposed a mathematical model for estimating EI (Thomas et al., 2010), which is derived from a differential equation approach for calculating weight dynamics (Thomas et al., 2011). The model requires inputs of age, gender, height at baseline and body weight throughout the observation period and incorporates expected fluctuations in RMR and PAEE in response to altered body mass. The predictive accuracy of the Thomas model was investigated in 23 subjects who were restricting caloric intake but maintaining physical activity levels. Estimates derived from the model were compared to EI values obtained by provision of food and secondly with the DLW and DEXA intake-balance method. Relative to food provision, errors were 41 ± 118 kcal/d for the initial 4 weeks and -22 ± 230 kcal/d weeks 4-12. Similar accuracy was observed relative to a gold-standard intake-balance method, with a maximum mean difference of -71 kcal/d, all of which were non-significant (Thomas et al., 2010). Despite these promising results, the Thomas model has had limited experimental application relative to an alternative model developed at the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK).

Hall and Chow showed that any change in body weight over a given interval can be used to derive estimates of change in EI (Chow & Hall, 2008; Hall & Chow, 2011). The parameters and assumptions of this model are complex and derived from a body of physiological research conducted over many decades, and these components are detailed in the methods of this thesis (**section 3.5.1**). In brief, given an initial body weight value, a differential equation can be derived which includes terms for the energetic cost of synthesis of FM and FFM ($\eta FFM$; 230 kcal/kg), the energy density of FM and

FFM and parameters to capture DIT and adaptive changes to the rate of EE associated with a change in body weight (Polidori et al., 2016). The model also includes a parameter which describes the relationship between changes in FM and FFM and is defined $\alpha$ = 10.4/FM (Chow & Hall, 2008) where the constant (10.4) controls the composition of weight change. Given a larger initial FM, the parameter $\alpha$ goes towards 0 and therefore the model predicts that those with a higher initial FM will partition a higher proportion of a given energy imbalance to FM (Forbes, 1987). Hall altered the Forbes equation to approximate the non-linear changes in FM which is thought to be more applicable in large weight losses (Hall, 2007). The model also has terms relating EE to body weight as determined by the quantity of FM and FFM. The parameter $\delta_0$ refers to PAEE estimates at baseline and is generally approximated as 10 kcal/kg/day (a physical activity level of ~1.6). The $\Delta\delta$ term describes changes in PAEE from baseline and it is assumed that $\Delta\delta$ =0, in the absence of objective or subjective estimates of activity (Guo et al., 2019). Based on these calculations the NIDDK model can be solved to estimate the $\Delta EI$ over an interval relative to the baseline maintenance requirement, where each interval is associated with an average weight and rate of change, determined by a linear regression over the weights in that interval (Sanghvi et al., 2015). To calculate $\Delta EI$ inputs of age, gender, weight and height are required, and the mathematical model is fully implemented in an accessible Java application (Guo, Personal communication). Notably, the primary outcome of the model is $\Delta EI$ relative to baseline requirements, rather than the more intuitive approach of absolute EI. The motivation for this is that estimating maintenance energy requirement is challenging without gold-standard measures. It is estimated that if no objective measures of TDEE are available, a 95% confidence interval may reach 400-500 kcals for baseline EI (Hall & Chow, 2011), which would substantially decrease the precision of the model.

The primary validation of the NIDDK model has been conducted in the CALERIE dataset, which is novel in that it has repeated body composition and DLW measures. Over 104 weeks in 140 subjects (minimum n=115 by week 104) the $\Delta EI$ estimates from the NIDDK model were compared to the intake-balance method (DLW and DEXA) at 4 time points (Sanghvi et al., 2015). Encouragingly, the model produced errors of ~40 kcal/d and root mean square deviation (RMSD) of 215 kcal/d, which was not statistically different from the criterion. Although the majority of the subject comparisons errors were < 132 kcal/day, limits of agreement reached ~ 1000 kcal/d implying that the NIDDK model has limited utility at the individual level. The

model also tended to underestimate the extent of the decrease in $EI$ during weight loss and the explanation for this may lie in the assumption of a constant rate of PAEE (Sanghvi et al., 2015).

After this validation, the NIDDK model has been applied in a range of cohorts. Combining the results of 15 obesity pharmacotherapy trials, the model was used to demonstrate 2-year trajectories of EI in placebo and drug-treated patients. In both groups, EI was calculated to reduce greatly at the beginning of treatment followed by a gradual return towards baseline (Göbel et al., 2014). In another application, patterns of EI were modelled in subjects treated with canagliflozin, a drug which leads to the excretion of glucose via urine, compared to a placebo. In this study, the treatment group was demonstrated to increase EI but decrease body weight owing to the loss of calories via urinary glucose excretion until a stable body weight was reached (Polidori et al., 2016). Next, the NIDDK model was used to estimate the magnitude of misreporting from 24-hour recalls in the DIETFITS trial over a year. A similar EI trajectory to the aforementioned studies was observed, where subjects initially reduced EI significantly (~800 kcal/d) and then exponentially returned to baseline by the final interval, day 360 (Guo et al., 2019). In contrast, self-reported EI was relatively constant, suggesting a greater degree of misreporting as weight relapsed. This study was novel as it incorporated self-reported PAEE data in a subset of participants, rather than assuming a constant rate of PAEE. This subset was observed to be within 70 kcal/day of the EI values derived from the assumed PAEE, implying a negligible difference with the incorporation of self-reported PAEE data (Guo et al., 2019).

A recognised limitation of these approaches is the lack of objective physical activity information and the assumption of a constant rate of PAEE (Polidori et al., 2016; Sanghvi et al., 2015). This assumption contributes to the limited individual-level predictive ability as the PAEE response to diet alterations is likely to be highly variable between subjects (Sanghvi et al., 2015). It is important to note that both the Thomas and NIDDK models have been validated in the CALERIE dataset, as their validation requires longitudinal body composition and DLW measures, which are uniquely offered by the CALERIE study. The CALERIE study was designed to assess caloric restriction in non-obese adults (Kraus et al., 2019) and the validity of these models in different groups and states of energy balance remains somewhat uncertain.

### 1.4.3 Body composition

If accurate and precise estimates of body composition are available, it is possible to quantify $\Delta ES$ with more objectivity and fewer assumptions than the aforementioned modelling approaches. The outcome measure of interest in the context of the present discussion is $\Delta FFM$ and $\Delta FM$, so a precise and valid measure, is needed. Many techniques are theoretically applicable to estimate $\Delta ES$, the subsequent discussion considers methodologies suited to and frequently used in energy balance research currently.

Bioelectrical impedance analysis (BIA) is a simple and accessible method with limited participant burden. It estimates FM, FFM and total body water through population-specific equations incorporating impedance values and anthropometric measurements (Mcguire & Ross, 2010). It is unlikely to be useful for individual-level estimates owing to its wide variability; one study reported a bias in FM of 0.8 kg, but wide variation, (2SD = 7.9 kg) relative to a four-compartment model (Jebb et al., 2000). Air displacement plethysmography (ADP) is a widely used, non-invasive technique which estimates body volume through the application of Boyle's Law, which describes volume and pressure relationships (Baracos et al., 2012). The measurement of body volume, together with body mass, permits calculation of body density (Fields et al., 2002) and subsequent estimation of FM% and FFM% with the models of Brozek or Siri (Brožek et al., 1963; Siri, 1961). The most common commercial ADP technology is the BodPod (Dempster & Aitkens, 1995). The BodPod shows good precision; one review found a mean within-subject coefficient of variation for FM% of less than 2.3% between measurements on different days (Fields et al., 2002). The same review compared the validity of the BodPod to hydrostatic weighing and reported a difference of <4% for FM% in all of the included 12 studies (Fields et al., 2002). It has been argued that variation between laboratories and testing conditions may explain the small variations observed (Fields et al., 2002) and repeated measures within laboratories may yield a better agreement. Thus, ADP using BodPod is a valid and reliable tool for body composition and therefore $\Delta ES$ in energy balance research. However, any two compartment model of body composition in the initial phase of weight loss, where fluctuations in glycogen, nitrogen and water are large, can introduce errors (Heymsfield et al., 2011).

Another potential approach to determine $\Delta ES$ is to estimate the total body water of a subject following the intake of a dose of labelled isotope and collection of biological samples to determine isotope quantity (Al-Ati et al.,

2015). The total body water method assumes water is maintained in consistent balance with FFM and as such, measurement of isotope dilution volumes facilitates the estimation of FFM (Duren et al., 2008). Estimates of total body water are converted to FFM based on the assumption of 73.2% hydration (Pace & Rathbun, 1945). This hydration factor holds amongst most populations with notable exceptions being pregnancy and oedema, where hydration will vary (Duren et al., 2008). A variable proportion of total body water is located in adipose tissue as extracellular fluid, therefore variability in the proportion of adiposity has the potential to reduce the accuracy of FFM estimates (Chumlea, 2006). Nevertheless, estimates give a precision (i.e. ±1 SD) close to 0.5 kg, corresponding to an error of ± 19.7 MJ in the energy content of FFM (Elia et al., 2003).

DEXA involves the administration of two low-energy x-rays and measures differences in attenuation through bodily tissues, permitting estimates of FM, FFM, bone mineral content and soft tissue for the whole body or specific regions (Mcguire & Ross, 2010). The measurement process takes less than 20 minutes and exposes the subject to tiny amounts of radiation, facilitating repeated measurements (Genton et al., 2002). The method shows excellent precision in short-term studies, with a coefficient of variation of <1% for determining the composition of bodily segments (i.e. limbs, trunk, etc) (Baracos et al., 2012). Yet, in a longer study of 7 days, de Jonge et al., (2007) reported an error of ± 300 g for FM, which would introduce a potential error of ± 2790 kcal in energy stored as fat. The authors also outline the potential for a large caloric deficit and the associated fluctuations in water balance to introduce error in DEXA measurements (de Jonge et al., 2007). The accuracy and the ability to measure repeatedly contribute to the popularity of DEXA for energy balance research (Ries et al., 2018; Shook et al., 2018) and as a gold-standard comparator for other methods (Heymsfield, 1997). It is however important to note some limitations such as the substantial cost of the scanners and potential inconsistencies between specific algorithms employed by manufacturers (Genton et al., 2002). Of the discussed body composition methods, ADP, TBW and DEXA all carry assumptions and these must be considered when estimating the $\Delta ES$. If these are not violated, these methods can offer an effective means of deriving individual-level estimates of $\Delta ES$, which is a necessity to estimate EI with intake-balance methods.

## 1.5 Conclusion

This chapter provided an introduction to the components of energy balance in humans as well as the potential for interaction between these components. Given the centrality of energy balance to an array of health and obesity-related fields there is a continued research interest in the ability to accurately and precisely estimate the EI, EE and ES of subjects. To ‘solve’ the energy balance equation and calculate the EI of a subject, these methods will need to include estimates of TDEE. Using currently available tools, it is not feasible to longitudinally estimate two of these three components. Mathematical modelling approaches show promise at the group level and are implementable in computer programs, but their individual-level accuracy is limited owing to the lack of objective estimates of TDEE. Wearable devices, which incorporate accelerometers and physiological sensors, offer a cost-effective means of capturing physical activities of people in free-living settings and when partnered with statistical learning techniques, they can theoretically be used to refine estimates of TDEE. Through wearable monitors and time-series body weight data, it is probably possible to provide objective and accurate estimates of changes in EI for large groups of subjects enrolled in weight maintenance trials.

# Chapter 2 – Aims and Objectives

This thesis aims to advance the measurement of energy balance in free-living subjects using wearable devices. A series of studies are conducted to develop and evaluate methodologies to provide continuous and accurate estimates of TDEE at the individual level. These estimates of TDEE will be incorporated into validated mathematical models of body weight dynamics to derive EI estimates for adults participating in a large, multi-centre weight loss maintenance trial (NoHoW trial). The aims of this thesis can be summarised as follows:

**Aim 1:** To investigate the validity of current wearable tracking technologies, against criterion measures for the estimation of heart rate and EE.

> Objective 1: To investigate the validity of current wrist and arm-worn devices for the prediction of EE compared to gold-standard methods in different activity modalities in previously published research (**chapter 4**).
>
> Objective 2: To compare the EE estimates for different sensor configurations utilised in previous devices /studies (**chapter 4**).
>
> Objective 3: To evaluate the minute-level accuracy of EE and heart rate estimates of the Fitbit Charge 2 (**chapter 5**) compared to indirect calorimetry and a Polar heart rate monitor, respectively.

**Aim 2:** To investigate methods to impute missing data in commercial activity monitors.

> Objective 4: To propose an imputation algorithm for imputing missing physical activity and EE data (NoHoW algorithm) and explore its validity relative to alternative imputation methods. Imputation bias is evaluated with different proportions of missing data, occurring at varying times of the day and days of the week (**chapter 6**).

**Aim 3:** To develop and validate machine learning algorithms for the prediction of EE.

> Objective 5: To evaluate machine learning algorithms for their ability to predict EE based on movement and physiological signals obtained from wearable devices, using indirect calorimetry as a reference standard (**chapter 7**).

Objective 6: To evaluate the predictive performance of the developed algorithms in a 14-day, free-living study (**chapter 8**).

**Aim 4:** To quantify EE, EI and energy balance in the NoHoW trial.

Objective 7: To utilise validated mathematical models, in combination with estimates of TDEE to provide a time-series of EI and EE for participants in the NoHoW trial (**chapter 9**).

# Chapter 3 – General Methods

## 3.1 Overview of projects

Three research studies contributed to this thesis and this chapter provides an overview of the methods employed throughout. The study descriptions in this chapter are not exhaustive and provide a general overview of the projects. The exact methods are clarified in the respective chapters.

### 3.1.1 Device validation study

The device validation study involved a laboratory study in which 59 adults were recruited. Body composition, anthropometric and RMR measures were made, and participants subsequently performed a standardised physical activity protocol, consisting of a series of submaximal activities including common household tasks, lying, sitting, walking, running, and cycling. Concurrently, EE was measured by a metabolic cart and simultaneous heart rate data and accelerometer counts were collected from wearable technologies (ActiGraph GT3-X, SenseWear Armband Mini, Fitbit Charge 2 and a Polar H7 chest strap). The study therefore has the aim of i) Validating the Fitbit Charge 2 and the SenseWear armband estimates of EE against indirect calorimetry (aim 1, objective 3) ii) Validating the heart rate estimates of the Fitbit charge 2 against a Polar heart rate chest strap (aim 1, objective 3) iii) Development of predictive models of EE (aim 3, objective 5). Further details of the activity protocol are provided in **chapter 7**.

### 3.1.2 TEED study

The total daily energy expenditure from wearables (TEED) study recruited healthy adult participants (n=30) for a two-part study. First, participants performed a submaximal exercise protocol with a more diverse set of activities than the device validation study. In the second part of the study, participants undertook a free-living component in which TDEE will be estimated over 14 days by the DLW method (detailed in **section 3.4.2.3**) and will ultimately serve as a free-living validation of a series of predictive models (detailed in **section 3.5.2**). Body composition measures at the start and end of the free-living period facilitated estimates of EI, in combination with TDEE estimates from each of the developed models. The aims of the TEED study were to understand the extent to which EE and EI can be estimated using inputs wearable devices in free-living humans.

### 3.1.3 NoHoW Study

The Horizon 2020 funded NoHoW project (Scott et al., 2019) (https://nohow.eu/) (ISRCTN88405328) included a randomised controlled trial testing the efficacy of an ICT-based toolkit to support WLM in European adults situated in the United Kingdom (Leeds), Denmark (Copenhagen), or Portugal (Lisbon). Participants were allocated to one of 4 arms after achieving ≥5% weight loss: (1) self-monitoring only (self-weighing and activity tracker), (2) self-regulation plus motivation, (3) emotion regulation, or (4) combined self-regulation, motivation, and emotion regulation. Participants were followed-up at 6, 12 and 18 months for a change in body weight, body composition, biomarkers, dietary intake, physical activity, sleep, and psychological factors, the primary outcome of body weight was measured at 12 months.

## 3.2 Ethics and recruitment

### 3.2.1 Device validation study

Recruitment for the device validation study was primarily from the Leeds centre of the NoHoW trial (discussed below) and an additional 15 participants were recruited from the University of Leeds and surrounding areas by recruitment posters, emails, and word of mouth. The device validation study was approved by the University of Leeds, School of Psychology ethics committee (PSC-407, 18th August 2018).

### 3.2.2 TEED study

The TEED study recruited participants by email invitation. Potential participants were identified as those that had expressed an interest in participating in future studies conducted at the University. Participants were also recruited within the University of Leeds research staff via word of mouth, email and recruitment posters. The TEED study was approved by the University of Leeds, School of Psychology ethics committee (PSC-744, 14th August 2019)

### 3.2.3 NoHoW Study

The NoHoW study was conducted between March 2017 and September 2019 at the participating institutions (detailed in **section 3.1.3**). Centre-specific recruitment strategies were adopted for 12 months (March 2017–March 2018) and included a commercial weight loss programme (UK, Slimming World); the Copenhagen Municipality weight management

services, Dieticians from the Danish Association for Dieticians and commercial slimming companies (e.g., Sense, Henrik Duer and Per Nielsen); registered clinical dieticians/nutritionists who provide weight management services in Lisbon; leisure centres; and local/national media coverage and advertisements. The trial was registered with the ISRCTN registry (ISRCTN88405328). Ethical approval was granted by each institutional ethics committee before study commencement; the Universities of Leeds (17–0082; 27 February 2017), Lisbon (17/2016; 20 February 2017) and Capital Region of Denmark (H-16030495, 8 March 2017). In total, 1627 participants (approximately balanced across centres) were enrolled. All participants provided informed consent before participation.

## 3.3 Inclusion and exclusion criteria

### 3.3.1 Device validation study

The device validation study included participants > 18 years of age, as determined by a screening questionnaire upon registration of interest. Exclusion criteria for the device validation study were:

- Medications associated with alteration to metabolic rate.
- The inability to ambulate without assistance.
- The presence or sign of cardiovascular, metabolic, renal disorders, illness or injury that provide an increased risk of medical events during low-to-moderate physical activity.

### 3.3.2 TEED study

The TEED study included participants > 18 years of age, as determined by a screening questionnaire upon registration of interest. Exclusion criteria for the TEED study were:

- Inability to attend HARU at required intervals.
- Diets not typical of a western diet which may influence respiratory quotient (i.e. very low calorie, ketogenic, high fat or high carbohydrate).
- Medications associated with alteration to metabolic rate.
- Inability to ambulate without assistance.

- The presence or sign of cardiovascular, metabolic, renal disorders, illness or injury that provide an increased risk of medical events during low-to-moderate physical activity.
- Participants without mobile phones compatible with the devices used in this study.

### 3.3.3 NoHoW study

To be included in the NoHoW trial participants must have been at least 18 years old, have a BMI (before weight loss) of ≥25 $kg/m^2$. Be able to verify ≥5% of weight loss in the last 12 months and remain 5% below their highest weight. The ability to use a smartphone and have access to a smartphone, tablet or computer with internet access and Wi-Fi. Ability to use standing scales for weight measurements and must not be over 150 kg. These were identified in a screening questionnaire and a subsequent phone screen.

Exclusion criteria for the NoHoW trial were:

- Inability to give informed consent.
- Lost weight through illness or surgical procedures.
- Pregnant or breastfeeding.
- Participation in another research intervention study that confounds with the aims of NoHoW (excluding local health interventions and weight management services).
- Inability to follow written material or telephone conversations in the language of the centre.
- Diagnosed with an eating disorder (e.g., anorexia nervosa, bulimia nervosa or purging disorder).
- Diagnosed with any condition that may interfere with increasing mild to moderate physical activities.
- Recent diagnosis with type 1 diabetes.
- Planned travel of more than 4 weeks.
- Living in the same household as an existing participant in the trial.

## 3.4 Physical and metabolic measurements

### 3.4.1 Anthropometric and physical measures

#### 3.4.1.1 Height

Height was measured to the nearest 0.1 cm with a Seca 704 s instrument (SECA, Germany). Measures were taken barefoot when participants were standing with their heels and back upright on the stadiometer.

#### 3.4.1.2 Waist and hip

Waist and hip circumference were measured in a private room to the nearest 0.1 cm in line with the navel and hip circumference was measured horizontally at the point of the greatest circumference of the hip.

#### 3.4.1.3 Blood pressure and resting heart rate

Systolic and diastolic blood pressure and resting heart rate were measured with a sphygmomanometer (Microlife BP A2 Basic, Gentle Technology, Microlife, Clearwater, FL, USA) when the participant was in a resting condition. The cuff was attached on the upper arm and three measures were conducted and averaged.

#### 3.4.1.4 Body mass and body mass index

Bodyweight was collected during the NoHoW trial to the nearest 0.1 kg using a SECA 704s instrument (SECA, Germany) with participants barefoot and wearing light clothing. For the device validation study and the TEED study, weight was obtained during the laboratory visits with the BodPod scales (discussed below). Body mass index (BMI) was calculated as follows*:*

$$\text{BMI } (\text{kg/m}^2) = \frac{\text{Body Mass}}{\text{Height}^2}$$

#### 3.4.1.5 Body composition

Body composition (2 compartment model) estimates were obtained whilst participants were wearing skin-tight clothing (i.e. swimming costume) and a swim cap. Measures were conducted with the BodPod (BodPod, Life Measurement, Inc., Concord, USA) which uses ADP to estimate body composition in two compartments. Before measures, the BodPod was calibrated for consistency and validity using a cylinder of known volume (50.03 L) and scales were calibrated regularly with two known weights (2 x 10kg), following the manufacturer instructions. The participant details

(height, age, gender and ethnicity) were entered into the software and then the participant was weighed using the calibrated BodPod scale. Participants were instructed to sit in the chamber and stay as still as possible but were instructed to breathe normally. Two measures were conducted and averaged if the measures agreed (difference < 150 ml), and if not a third measure was conducted and the three measures were averaged. The BodPod measures the displacement of air by the participant's body (thoracic gas volume is estimated) and body volume is calculated according to Boyle's law. The density of the participant's body is then estimated by dividing the mass by the volume. Body fat (%) is then estimated by the equation of Siri (Siri, 1961):

$$Body\ Fat\ (\%) = \left(\frac{4.95}{Density - 4.5}\right) \times 100$$

### 3.4.2 Energy expenditure

#### 3.4.2.1 The principles of indirect calorimetry

Indirect calorimetric measures were used throughout this thesis and each of these is described below. The basis of indirect calorimetry is that all of the biological processes, in which an organic substrate is oxidised and energy is produced, require the consumption of $O_2$ and the production of both $CO_2$ and $H_2O$. Therefore, the measurement of $O_2$ consumption $(\mathrm{V}\mathrm{O}_2)$ and/or $CO_2$ production $(\mathrm{V}\mathrm{CO}_2)$ can be used to infer the heat production (EE) associated with these processes (Elia & Livesey, 1992). When the term 'energy production' is used, it is in reference to the metabolic process whereby adenosine triphosphate (ATP) is produced from the free energy of dietary nutrients (Ferrannini, 1988). Indirect calorimetry measures $\mathrm{V}\mathrm{O}_2$ and $\mathrm{V}\mathrm{CO}_2$ to infer the heat production of the body. Specifically, when any of the predominant energy sources (carbohydrate, protein or fat) are metabolised by bomb calorimetry, the amount of $O_2$ consumed and $CO_2$ produced differs, which gives a different respiratory quotient (RQ; where $\mathrm{RQ} = \mathrm{CO}_2/\mathrm{O}_2$) per macronutrient. Slight differences exist in the specific heat equivalents, $O_2$ equivalents and RQ of macronutrients in the published literature (Ferrannini, 1988; Livesey & Elia, 1988; Weir, 1949) although in practice these produce tiny differences in EE estimates stated (Montoye et al., 1996. pp. 15 - 21). Using $O_2$ equivalents (i.e. heat produced per litre of $O_2$), it is possible to solve a system of standard stoichiometric equations (for specific equations see (Ferrannini, 1988)) to obtain an estimate of the EE rate, given a

measure of $VO_2$, $VCO_2$ and nitrogen excretion (Ferrannini, 1988). An error of 100% in the nitrogen excretion value is thought to lead to an EE bias of ~1% and therefore when subjects are consuming mixed diets, it may be excluded from heat production equations (Montoye et al., 1996; Weir, 1949).

The general assumptions of indirect calorimetry are as follows (McLean & Tobin, 1988; Mtaweh et al., 2018):

1. The metabolite ultimately results in heat/energy production (i.e. oxidation of nutrients is complete).
2. The combustion or synthesis of the dietary macronutrients is the end result of all the biochemical reactions occurring in the body.
3. The oxidation of glucose, fat, or protein results in a specific RQ.
4. The loss of substrates to faeces and urine is minimal.

These assumptions are not violated in healthy subjects and extremely close agreement between indirect calorimetry and direct calorimetry has been shown on numerous occasions since the early 19$^{th}$ century (McArdle et al., 2015; McLean & Tobin, 1988). In the open indirect calorimetry systems used in this thesis, the ventilation rate and the content of the inspired air are measured, and subsequently, $VO_2$, $VCO_2$, RQ and the rate of EE are calculated. Thus, the accuracy of indirect calorimetry is largely dependent on the system-specific sensors (Montoye et al., 1996; Mtaweh et al., 2018).

#### 3.4.2.2 Resting metabolic rate

Resting metabolic rate was estimated using the GEM indirect calorimeter ((GEM, NutrEn Technology Ltd, Cheshire, UK), an open circuit stationary indirect calorimetry system which uses a ventilated canopy. The transparent canopy hood is placed over the subject's head such that air is drawn through a Nafion tube. Next, the exhaled air mixes with ambient air in a chamber, where $VCO_2$ and $VO_2$ can then be estimated (Kennedy et al., 2014). Measures were conducted in the early morning, following an overnight fast and before the participation in physical activity. Before each measure, the GEM was calibrated against reference gases, which was conducted over approximately 10 minutes, following manufacturer instructions whilst the participant lay flat. The RMR was estimated whilst the participant lay supine (and without talking, moving or falling asleep) for 30 minutes. The estimation of RMR is obtained by a 5-minute method (Sanchez-Delgado et al., 2018); RMR was calculated by removing the first 5-minute interval and selecting the 5-minute interval in which the coefficient of variation was the lowest across

$VCO_2$, $VO_2$ and respiratory exchange ratio. The RMR is then estimated by the GEM software by the modified Wier equation (Weir, 1949):

$$RMR = (3.94 \times VO_2) + (1.11 \times VCO_2)$$

### 3.4.2.3 Exercise energy expenditure

In the laboratory exercise studies, $\dot{V}O_2$, $\dot{V}CO_2$ and EE were derived during activities with a stationary metabolic cart: Vyntus CPX, (Jaeger-CareFusion). Breath-by-breath $\dot{V}O_2$ and $\dot{V}CO_2$ were collected with a facemask, which was connected to the Vyntus system. The accuracy with which such systems can estimate breath-by-breath EE is highly dependent on the sampling time delay, as a slowed delay time can result in errors in $\dot{V}O_2$ at high rates of ventilation (Overstreet et al., 2017). The Vyntus CPX has an extremely short delay time for $\dot{V}O_2$ and $\dot{V}CO_2$ (Perez-Suarez et al., 2018). The Vyntus is, however, a relatively new system and the majority of the relevant experimental validation studies have been conducted in the predecessor, the JAEGER Oxycon Pro, which was demonstrated to measure at < 1.1 % error for $\dot{V}O_2$ compared to a Douglas bag criterion for a large range of ventilatory rates in cycling activity in elite athletes (Foss & Hallén, 2005). Perez-Suarez and colleagues conducted comprehensive butane experiments, simulating low, moderate and intense exercise at 0.8, 1.3, and 6.4 L/min$^{-1}$ $\dot{V}O_2$. They showed that the RQ deviated by <1.5% for all comparisons and subsequently stated that the Vyntus is "*exceptionally accurate and precise for measuring the stoichiometric RQ of butane combustion*" (Perez-Suarez et al., 2018). Before each measurement, the Vyntus was calibrated automatically for volume and gas relative to a reference. Breath by breath EE data was calculated by the system, (assuming a minimal contribution of protein oxidation) (Péronnet & Massicotte, 1991) which was aggregated to the minute level and used as the outcome variable in **chapter 5**. Rather than absolute EE (kcal), the machine learning models trained in **chapter 7** used metabolic equivalents (METs) as the outcome variable, which is defined as the minute-level EE as a multiple of each participant's minute-level RMR.

### 3.4.2.4 Doubly labelled water

At the time of submission, the analysis of the DLW samples collected in this thesis is not complete. Data collection was completed by March of 2020 in anticipation of submitting this thesis in February 2021, but university closures, lockdown laws and logistical issues have prevented this from being

completed. It is anticipated that the affected chapters (**chapters 8 and 9**) will be published in an academic journal after this return of this result. It is likely that these models will be continuously refined in light of new evidence.

The criterion method of TDEE in this thesis was the DLW method, which can be considered to be a form of indirect calorimetry as heat production is estimated by indirect means. The DLW method is based on calorimetric principles to estimate the TDEE based on the energy equivalents of $CO_2$. $CO_2$ production is estimated from isotope elimination rates of two stable isotopes, deuterium ($^2H$) and oxygen 18 ($^{18}O$). The method involves the ingestion of $^2H$ and $^{18}O$ (typically orally). The body water of the subject becomes enriched with the isotopes and over the course of the measurement period, the difference in washout kinetics for each isotope is determined from isotopic enrichments of urine samples taken at regular intervals after isotopic equilibration in body pools (Westerterp, 2017). The $^2H$ isotope only labels the body water pool and is excreted as $H_2O$ only. $^{18}O$ exchanges with $CO_2$ in the body's bicarbonate pools, equilibrates with the body's $CO_2$ and $H_2O$ pools and as such is lost as $CO_2$ and $H_2O$. This means that $^{18}O$ is lost at a higher rate than $^2H$, and the difference in elimination corresponds to the $CO_2$ production (Westerterp, 2017). Several assumptions and potential sources of error for the DLW method must be stated (Montoye et al., 1996. pp 15-21; Speakman, 1997; Speakman, 1998):

1. The number of water molecules in the body is constant (i.e. there are no large fluctuations in hydration during a period of measurement)
2. There is no exchange of $^2H$ or $^{18}O$ with nonaqueous bodily tissues
3. $^2H$ and $^{18}O$ only leave the body via $H_2O$ or $CO_2$
4. Turnover rates of the isotopes are constant for the duration of the measurement
5. The method of sampling (i.e. urine or saliva) is representative of the TBW
6. The isotopes once excreted do not re-enter the body
7. The RQ of the diet can be estimated with accuracy
8. The abundance of isotopes in the background sample is typical of the true values

For assumption 1, the body water must not fluctuate markedly although it has been suggested that this error would be negligible if the water pool remains within ± 10% (Nagy, 1980). Concerning assumption 2, associations

between TBW for each of the isotopes has been reported at r = 0.998 with a difference of ~1% (Schoeller & Van Santen, 1982), so this is not likely to introduce large errors. Assumption 3 presents as an issue if the isotopes are lost disproportionally (i.e. through alternative excretion pathways), though the associated error for this is thought to be in the region of 2% (Montoye et al., 1996. pp 15-21). For assumption 4, fitting a regression model to the observed data is likely an effective strategy to overcome any issues with variance (Cole & Coward, 1992) and for assumption 7, estimating or measuring RQ in people consuming mixed diets should not produce an error above 2% (Black et al., 1986). Recent evidence suggests that ketogenic diets may bias the measure, although the errors associated with these dietary practices are small and are likely to be smaller than the precision expected with modern analytical technology (Hall et al., 2019). With these, and other analytical considerations in mind, it was stated that the theoretical coefficient of variation for DLW is between 4 and 8% (Schoeller, 1983), and this is thought to still be the case (Westerterp, 2017). Extensive consideration has been given to these sources of error elsewhere (Cole & Coward, 1992; Speakman, 1997).

The use of the method in the present thesis was as follows:

Participants provided a background urine sample upon arrival at the laboratory for visit 2 of the TEED study, which was not the first void of the day. Baseline samples were labelled with the time, date and identification and information was stored in a locked spreadsheet. Next, participants consumed a bodyweight-specific dose of $^{2}H$ and $^{18}O$ and the exact time and date of consumption were recorded, a sample of the dose was also retained at the University of Aberdeen for analysis (see below). Participants consumed the entire dose under the supervision of a researcher and each vessel was swilled with water twice to collect any remaining drops. Participants were required to provide their initial sample 6-8 hours after the dose and to ensure this was not missed researchers called the participant and alarms were set on their phone. Participants were then instructed to collect samples every 48 hours, which must not have been the first void of the day. Participants were required to seal, label (identification, time, date) and freeze samples and store them in provided bags. Participants returned all samples to the laboratory on their third visit of the TEED study. All samples were kept frozen until analysis. Analysis of the isotopic enrichment of urine was performed blind, using a Liquid Isotope Water Analyser (Los Gatos Research, USA) (Berman et al., 2012). Initially, the urine was vacuum

distilled, and the resulting distillate was used for analysis. Samples were run alongside five lab standards for each isotope and International standards to correct delta values to ppm. Daily isotope enrichments were $\log_e$ converted and the elimination constants ($k_o$ and $k_d$) were calculated by fitting a least-squares regression model to the $\log_e$ converted data. The back extrapolated intercept was used to calculate the isotope dilution spaces ($N_o$ and $N_d$). A two-pool model, specifically equation A6 from Schoeller et al (Schoeller et al., 1986) as modified by Schoeller (Schoeller, 1988) was used to calculate rates of $CO_2$ production as recommended for use in humans (Speakman, 1993).

### 3.4.3 Digital tracking technologies

All of the experimental studies conducted in this thesis utilised wearable devices or digital tracking technologies and each of these is discussed below. Inclusion criteria and data processing for each of the analyses is further specified in the respective chapters.

#### 3.4.3.1 Fitbit Charge 2

The FB (Fitbit Inc., San Francisco, CA, USA) is a wrist-worn activity monitor which estimates heart rate, steps, EE and physical activity, based on data obtained from incorporated sensors via proprietary algorithms. Acceleration is measured in three axes and heart rate estimates are obtained through a patented technology called ‘PurePulse’, which uses light-emitting diodes on the surface of the skin to monitor blood volume continuously (Benedetto et al., 2018). Data are aggregated to the minute-level and synced via the Fitbit mobile application to Fitbit servers through an application programming interface. The device was fitted a finger’s width above the non-dominant wrist and was configured with participant weight, height, sex and date of birth.

#### 3.4.3.2 SenseWear armband

The SWA (BodyMedia Inc., Pittsburgh, PA) is a small, non-invasive activity monitor which is worn on the upper arm and estimates tri-axial accelerometry, galvanic skin response, skin temperature and heat flux at the minute level. The SWA is a research focussed model and represents one of the most recent iterations from these devices, progressing from bi-axial models previously (Reeve et al., 2014). These outputs are processed by proprietary algorithms to derive variables of interest (i.e. EE, minutes spent in activity categories, steps, sleep etc). Data were downloaded and processed using the SenseWear® Pro 8.0 software, algorithm v5.2. The

SWA was fitted with an elastic strap around the non-dominant arm and initialised using participant weight, height, sex, date of birth and smoking status and the software estimates the participants RMR using a world health organisation (WHO) equation.

#### 3.4.3.3 Actigraph GT3-x & GT9-x

An Actigraph GT3-x accelerometer (AG; ActiGraph Corp., Pensacola, FL) measured acceleration along vertical, horizontal and perpendicular axes at a sample rate of 30 Hz. In the TEED study, participants wore an ActiGraph GT9-x Link accelerometer, which uses the exact same accelerometer, sampling and filtering methods as the GT3-x in the device validation study. The GT9-x differs in terms of size and also features bluetooth connectivity, which allows continuous integration of heart rate data from a polar heart rate strap. The actigraph models were always worn on the non-dominant wrist. Accelerometer data were downloaded and features were extracted at the minute-level using the feature extraction tool within the ActiLife software (Version 6.11.9).

#### 3.4.3.4 Polar heart rate

Heart rate was assessed during the laboratory protocols using a Polar m400 Monitor Watch (Polar Electro, Kempele, Finland) and a Polar H7 chest strap (Polar Electro, Kempele, Finland), which transmitted second-level data via a Bluetooth connection. Data were uploaded to the Polar flow online application, then downloaded and aggregated to minute-level for analysis. The Polar H7 was used in the device validation study and it has been shown to have a near-perfect correlation with an electrocardiogram during many exercise modalities (Gillinov et al., 2017). For the TEED study, a Polar H10 heart rate sensor was used and data were obtained via the ActiLife software (Version 6.11.9), as the monitor transmits heart rate data via a bluetooth connection directly to the Actigraph.

#### 3.4.3.5 Aria scales

In the NoHoW study, participants were provided with a Fitbit aria scale, which was used to obtain body weights from participants between clinical investigation days. Data were synced via the Fitbit mobile application to Fitbit servers through an application programming interface and subsequently obtained by the NoHoW datahub from the Fitbit API which was run by the James Hutton Institute. In a previous validation study, the Fitbit Aria as been shown to be highly accurate, relative to research-grade scales across a range of weights (Shaffer et al., 2014). All participants were

required to weigh themselves twice per week, after emptying their bladder and whilst wearing light/no clothing. Data were also screened for outliers by removing any instance in which weight deviated from the first observation by ± 5% in a single week (Turicchi, O'Driscoll, Horgan, Duarte, Palmeira, et al., 2020). Next, to obtain a daily weight value, linear interpolation was used to fill gaps in between bodyweight measurements, where the gap was less than 182 days which is equivalent to one body weight every clinical investigation day in the NoHoW trial. Body weight data were then 'smoothed' using a locally estimated smoothing regression model. The purpose of this is to smooth through small fluctuations which are unlikely to be due to energy imbalance and more likely reflect slight differences in weighing conditions (clothing, nutritional status, hydration). This was conducted with the Python module "Statsmodels" (Seabold & Perktold, 2010).

## 3.5 Modelling approaches and statistical methods

### 3.5.1 A mathematical model of energy intake

A linearised model was used to approximate the change in EI in the study reported in chapter 9. The model has previously been validated (Sanghvi et al., 2015) and has been applied in experimental settings (Guo et al., 2019; Polidori et al., 2016). The model is used to approximate change in Δ$EI$ over a predefined interval $i$ relative to the rate of EI required for baseline weight maintenance. The parameters are detailed below (table 3.1):

$$\Delta EI_i = \rho \frac{dBW_i}{dt} + \varepsilon_i\left(\overline{BW}_i - BW_0\right) + \frac{\Delta\delta_i}{1-\beta} BW_0$$

Here, $\rho$ is the energy density associated with the change in body weight, *BW,* and is defined:

$$\rho = \frac{\eta_{FM} + \rho_{FM} + \alpha\eta_{FFM} + \alpha\rho_{FFM}}{(1-\beta)(1+\alpha)}$$

and $\varepsilon_i$ defines how EE depends on *BW*:

$$\varepsilon_i = \frac{1}{(1-\beta)}\left[\frac{\gamma_{FM} + \alpha\gamma_{FFM}}{(1+\alpha)} + \delta_0 + \Delta\delta_i\right]$$

The parameters $\gamma_{FFM}$ and $\gamma_{FM}$ are the EE coefficients for FFM and FM, respectively. Parameters $\rho_{FM}$ and $\rho_{FFM}$ describe the energy densities of changes to FM and FFM. The parameter $\delta_0$ represents the PAEE at baseline which was set to the mean of the observations of the first two weeks of the study. The parameter $\Delta\delta_i$ is the change in PAEE for the ith interval relative to baseline, both PAEE parameters are measured in kcal/kg/day. Parameters $\eta_{FM}$ and $\eta_{FFM}$ account for the energetic cost of tissue deposition and the parameter $\alpha$ describes the relationship between changes of FFM and FM, $\alpha = dFFM/dFM = CFM$ and the parameter $C$ = 10.4 kg is the Forbes parameter (Forbes, 2000). The interval specific change of mean body weight versus baseline over each interval is denoted $\overline{BW}_i - BW_0$, and the moving average of the measured body weight time course is used to estimate the change in body weight per interval, $dBW_i/dt$. A time interval of 28 days was used for analyses which provides a more granular analysis than previous uses of the model with self-report PAEE measures (Guo et al., 2019), but would be less likely to be subject to errors associated with short term weight fluctuations (Bhutani et al., 2017). The model was implemented using a Java application developed by researchers at the NIDDK. Table 3.1 shows the numeric terms of the model (Sanghvi et al., 2015).

**Table 3.1** Parameters of the mathematical model.

| Parameter | Value | Summary |
|---|---|---|
| $\delta_0$ | Average of activity in the first two weeks of the study (kcal/kg/d) | Estimated physical activity at baseline |
| $\Delta\delta_i$ | Difference between $\delta_0$ and the average of available days in the interval (kcal/kg/d) | Physical activity changes in the interval |
| $\rho_{FM}$ | 9300 kcal/kg | Assumed energy density of fat mass |
| $\rho_{FFM}$ | 1100 kcal/kg | Assumed energy density of fat-free mass |
| $\gamma_{FM}$ | 3.2 kcal/kg/d | Estimated caloric expenditure rate of fat mass |
| $\gamma_{FFM}$ | 22 kcal/kg/d | Estimated caloric expenditure rate of fat-free mass |
| $\eta_{FM}$ | 180 kcal/kg | Assumed caloric cost of fat synthesis |
| $\eta_{FFM}$ | 230 kcal/kg | Assumed caloric cost of protein synthesis |
| $\beta$ | 0.24 | Assumed dietary and adaptive thermogenesis |

### 3.5.2 Predictive algorithms

**Chapter 7** involves the development of several algorithms, some of which are applied in **chapter 8**. For each of the algorithms, tuning experiments were conducted to identify the optimal hyperparameters of the model and the methods and results of this analysis are shown in **chapter 7**. A general summary of the algorithms used in this thesis follows below:

#### 3.5.2.1 Random forest

For regression and classification tasks the random forest algorithm was used (Breiman, 2001). The random forest is a powerful machine learning algorithm which is a generalisation of the 'bagging' technique. In bagging, the general premise is that the average of a large number of noisy decision trees serves to produce an ensemble estimator with low bias. Random forests train multiple decision trees on subsamples of the data and importantly when splitting these decision trees only a subsample of the potential predictors is used, which serves to 'decorrelate' the trees. The predictions of each tree can then be combined to produce a majority vote (classification) or a continuous prediction (regression) (Hastie et al., 2009. pp 587-601). The optimal hyperparameters of the algorithm were estimated in the tuning experiments and included: the number of trees (B), number of samples required to split a tree, number of samples per leaf, total predictors and the depth of trees. In regression, the quality of a split was assessed with mean squared error and in classification, Gini impurity was used. Algorithms were implemented using the 'RandomForestClassifier' and 'RandomForestRegressor' classes in Scikit Learn (Pedregosa et al., 2011). A general algorithm of the random forest is shown in algorithm 3.1.

for b=1 to *B*:

> Draw a bootstrapped sample Z* of size N from the training data
>
> Train a tree $T_b$ on this bootstrapped sample. To train this tree, repeat the following steps until the minimum node size is reached
>
> > Select *m* variables at random from the *p* predictors
> >
> > Select the best variable and split-point amongst the *m* variables
> >
> > Split the node into daughter nodes

Output the ensemble model $\{\mathbf{T_b}\}_1^{\mathbf{B}}$

To make a regression prediction:

$$\hat{\mathbf{f}}_{\mathbf{rf}}^{\mathbf{B}}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{B}}\sum_{\mathbf{b=1}}^{\mathbf{B}} \mathbf{T_b}(\mathbf{x})$$

And to make a classification prediction:

$$\hat{\mathbf{C}}_{\mathbf{rf}}^{\mathbf{B}}(\mathbf{x}) = \mathbf{majority\ vote}\left\{\hat{\mathbf{C}}_{\mathbf{b}}(\mathbf{x})\right\}_1^{\mathbf{B}}$$

**Algorithm 3.1.** A representation of the random forest algorithm. Adapted from (Hastie et al., 2009. pp 587-601).

#### 3.5.2.2 Gradient boosting

For regression and classification tasks the gradient boosting algorithm was employed. Like random forests, this algorithm is a tree-based ensemble method. However, where random forests may be considered to use a 'bagging' approach, gradient boosting uses 'boosting' to make predictions. A general algorithm for gradient boost regression is shown in algorithm 3.2. For classification approaches, a similar but more complex approach is taken, where one tree is grown per class, which can be used to probabilistically assign the most likely class (Hastie et al., 2009. pp 337-384). Boosting involves growing small (weak) decision trees sequentially and does not involve bootstrapping. Each tree is trained using residuals (r) of the previous estimator and subsequently added into the fitted function to update the residuals. In the boosting phase, a learning rate/ shrinkage parameter $\lambda$ penalises the contribution of each tree to the overall model, thereby slowing the learning (Hastie et al., 2009. pp 337-384). The trees grown can be extremely small and this size is controlled by the parameter *d* in the algorithm. Gradient boosting hyperparameters were tuned in the random search experiments and included the number of boosting stages, the learning rate, the number of samples required to split a node and number of

samples per leaf. In regression, the loss function was least squares and in classification, deviance was used. Algorithms were implemented using the 'GradientBoostingClassifier' and 'GradientBoostingRegressor' classes in Scikit Learn (Pedregosa et al., 2011).

Set $\hat{\mathbf{f}}(x) = 0$ and $\mathbf{r_i} = \mathbf{y_i}$ for each observation

for b=1 to B:

Fit a tree $\hat{f}^b$ with d splits and (d + 1 terminal nodes) to the training data

Update $\hat{f}$ by adding a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

Update the residuals:

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

Output the model:

$$\hat{\mathbf{f}}(\mathbf{x}) = \sum_{\mathbf{b=1}}^{\mathbf{B}} \boldsymbol{\lambda} \hat{\mathbf{f}}^{\mathbf{b}}(\mathbf{x})$$

**Algorithm 3.2.** A representation of the gradient boosting algorithm for Regression. Adapted from (James et al., 2013. pp 321-324).

### 3.5.2.3 Neural networks

The third algorithm used in both regression and classification tasks was artificial neural networks. Neural networks allow for complex, non-linear functions to be modelled and are comprised of layers of interconnected 'neurons', which may be compared to biological neurons. Many 'hidden units' or 'neurons' serve as unobserved variables, which are linear combinations of the input variables. At each hidden neuron, inputs are subjected to a numerical activation function and then passed through hidden layers of neurons to an output layer (Kuhn & Johnson, 2013. pp 141 & 333).

$$h_k(\mathbf{x}) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \beta_{jk}\right)$$

Here, the linear function is subject to an activation function $g()$. In the above equation, $\beta_{jk}$ represents the *j*th input variable of the *k*th neuron (Kuhn & Johnson, 2013). The output of this first hidden layer is then passed to

another hidden layer or to an output layer, where the output is modelled by another activation function to produce a prediction. In the training process, the inter-neuronal weights of the network are refined relative to a loss function (i.e. mean squared error or cross-entropy). Neural networks in the classification studies sought to minimise the sparse categorical cross-entropy and in the regression setting the loss was mean squared error. The learning rate of each network, the number of layers and the number of neurons were all selected based on the results of a randomised search, which is detailed in **chapter 7**. Regression neural networks used the 'relu' activation function in the hidden layers and classification models used a 'softmax' activation in the output layer, both classification and regression networks used the Adam optimiser.

#### 3.5.2.4 K Nearest Neighbors

For classification tasks, the k-nearest neighbors (KNN) algorithm was used. This algorithm assigns a given point to a particular class based on the majority class of the k-nearest neighbors, where the neighbors of a given point are defined by a distance metric (i.e. Euclidian, Minkowski or Manhattan). Hyperparameters adjusted in the training process included the number of neighbours in each neighbourhood (k), distance metrics and the weight applied to each of the observations in a neighbourhood. The KNN algorithm was implemented with Scikit learn (Pedregosa et al., 2011), using the 'KNeighborsClassifier' class.

#### 3.5.2.5 Support vector machine

The final classification model tested was the support vector machine classifier with the Radial Basis Function (Kuhn & Johnson, 2013. pp 343). A support vector machine aims to find a separating hyperplane between classes by maximising the distance between the points and the hyperplane. In **chapter 7**, the cost of misclassification of points in training (c) and 'gamma' which defines the magnitude of the effect of specific training examples, were tuned in randomised search experiments. The support vector machine classifier was implemented with the 'SVC' class in Scikit Learn (Pedregosa et al., 2011).

### 3.5.3 Computational methods

#### 3.5.3.1 Datahub

A NoHoW data-hub was developed and maintained by the James Hutton Institute (Edinburgh). The data-hub is a data architecture with the role of collating, monitoring and storing the data collected at clinical investigation

days, or from the digital tracking technologies. Data from each centre were entered into trial management software (Easy Trial: www.easytrial.net) and researchers at each centre conducted quality and consistency checks.

#### 3.5.3.2 Computing hardware

The simulation analyses conducted in this thesis were undertaken on ARC3, part of the High-Performance Computing cluster at the University of Leeds, UK. ARC3 is a Linux- based system using the CentOS 7 distribution. Most of the analyses conducted in this thesis were performed locally on a Windows-based computer with an intel i7-8750H with 32GB RAM and 12 logical processors or more frequently, a Windows-based computer with an Intel i9-9900K with 64GB RAM, 16 logical processors and an NVIDIA GeForce RTX 2080 Super graphics processing unit. The graphics processing unit in the PC facilitates the training of the neural network models utilised in **chapter 7** and **8**.

### 3.5.4 Statistical analysis

#### 3.5.4.1 Validation methods

Root mean squared error (RMSE) is reported throughout this thesis and was defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

and mean absolute percentage error (MAPE), which is defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

These methods were implemented with the metrics package in R (Hamner et al., 2018). In RMSE and MAPE, the predicted value is denoted $\hat{y}_i$ and the reference value is $y_i$. Equivalence tests (Lakens et al., 2018) were employed to determine if the true and predicted values were statistically equivalent. The tests used equivalence bounds of ± 10% and to be considered equivalent the 90% confidence interval must fall within the equivalence bounds (Lee et al., 2014), and all equivalence tests were conducted with the 'TOSTER' package in R..

For classification tasks, the Kappa statistic was employed, which compares the accuracy of the predictions to that of a random system. Also, accuracy, where accuracy is the proportion of the cases that were classified correctly and the F1-Score, defined as:

$$\mathrm{F1-score} = 2 \cdot \frac{(\mathrm{precision} \cdot \mathrm{recall})}{(\mathrm{precision} + \mathrm{recall})}$$

Where precision and recall are defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

Where $\mathrm{TP} = \mathrm{True\ positive}$, $\mathrm{FP} = \mathrm{False\ Positive}$, $\mathrm{FP} = \mathrm{False\ positive}$, $\mathrm{FN} = \mathrm{False\ negative}$. Classification statistics were calculated with the Caret package in R (Kuhn, 2008).

#### 3.5.4.2 General statistical reporting

Unless otherwise stated, data are presented as means ± standard deviation. Statistical analyses, visualisations and data processing steps were conducted in Python (van Rossum & Drake, 2009), within a Jupyter notebook or Jupyter Lab environment and R, within an RStudio environment and the specific versions are stated within the chapters. A wide range of packages, modules and statistical methods were utilised and these are referenced where necessary in specific chapters. A p-value of $<0.05$ is used to determine statistical significance where p-values are reported.

# Chapter 4 – A meta-analysis of the validity of activity monitors for the measurement of energy expenditure

## 4.1 Introduction

The first chapter of this thesis highlighted the increasing prevalence of obesity around the world (Ells et al., 2018) and the concerning projection that by 2050, 60% of males and 50% of females may be obese (Agha & Agha, 2017). The physiological, psychological and environmental factors which result in a chronic imbalance between EI and EE (and therefore a weight increase) must be studied to facilitate the development of effective interventions and treatments. However, to comprehensively and precisely map these relationships, and improve behaviour change interventions themselves, accurate, objective measures of energy balance behaviours are required.

The DLW method (See **section 1.3.3.2** and **3.4.2.3**) is considered the gold standard for the measurement of free-living EE (Seale et al., 1993); however, the considerable costs and analytical requirements of the method limit its feasibility in large cohort studies (Delany, 2012). Indirect calorimetry methods (See **section 1.3.3.1** and **3.4.2**) represent the most commonly employed criterion measures for the assessment of the energy cost of activities but are limited to structured protocols, usually within a laboratory (Hills et al., 2014).

Wearable devices which use triaxial accelerometry to derive an estimate of EE have been available for research purposes for some time (Lyden et al., 2011). These devices are worn on the hip, thigh or lower back, as proximity to the centre of mass is thought to more accurately reflect the energy cost of movement (Chen et al., 2003). However, participant comfort and compliance is a recognised issue (Diaz, Krupka, Chang, Shaffer, et al., 2016) and therefore traditional wearable devices have limited long-term, free-living measurement capability. The use of wrist-worn activity monitors by both consumers and researchers has dramatically increased (Wright et al., 2017) facilitated by improved battery longevity and miniaturisation of hardware required to produce interpretable data (Shcherbina et al., 2017). Recent consumer devices include triaxial accelerometers, heat sensors and photoplethysmography heart rate sensors (Woodman et al., 2017). This information can potentially be incorporated into predictive models to improve

the estimation of EE relative to accelerometry alone (Silva et al., 2015). Though evidence suggests that the accuracy of incorporated sensors and the derived activity metrics within commercial activity monitors, when compared with criterion measures, is variable (Evenson et al., 2015; Stahl et al., 2016) and may vary with the type and intensity of activity (Koehler & Drenowatz, 2017).

### 4.1.1 Chapter aims

Given the recent popularity of wrist and arm-worn activity monitors, it is critical to determine their validity for the estimation of EE (Evenson et al., 2015). The meta-analysis conducted in this chapter aimed to investigate the accuracy of EE estimates from wrist or arm-worn devices in different activities. Secondary aims were to investigate the importance of specific sensors within devices and to compare commercial and research-grade devices for their accuracy in estimating EE relative to criterion measures. It was hypothesised that the inclusion of physiological sensors in addition to accelerometry would provide a more accurate estimate of EE (Brage et al., 2015). Further, it was hypothesised that the performance of research-grade devices would be superior to commercial devices.

## 4.2 Methods

This systematic review and meta-analysis adhered to PRISMA diagnostic test accuracy guidelines (McInnes et al., 2018) (See online supplementary material 1 (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020)) and was prospectively registered in the PROSPERO database (CRD42018085016).

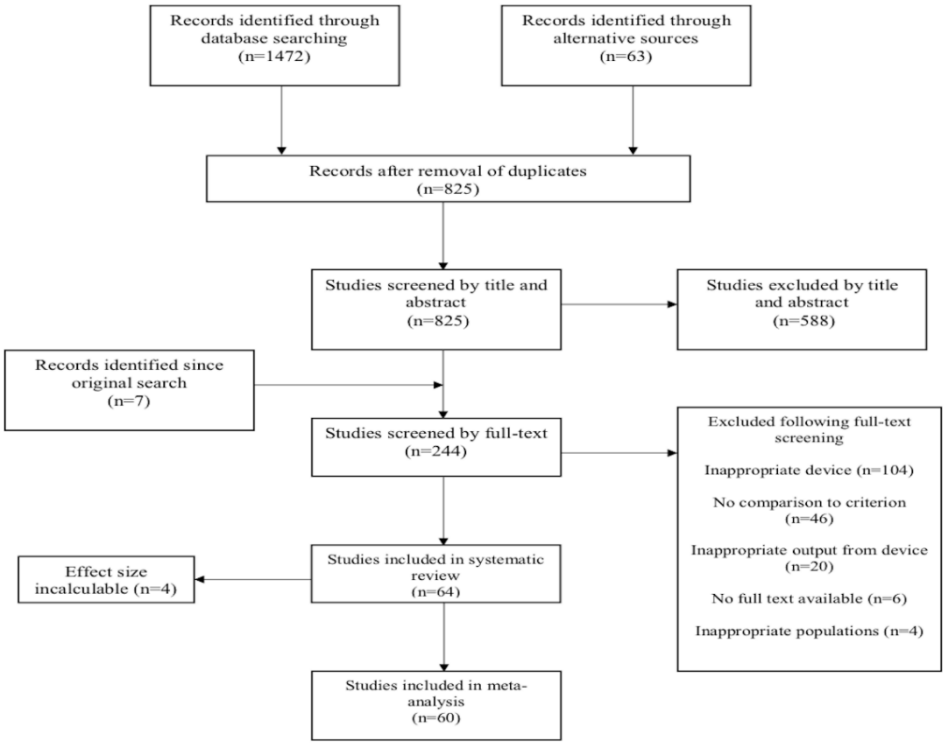



**Figure 4.1** A flow diagram of the study selection for the meta-analysis.

### 4.2.1 Search strategy

SportDISCUS (EBSCOHost), PubMed, Medline (Ovid), PsycINFO (EBSCOHost), EMBASE (Ovid) and CINAHL (EBSCOHost) were searched for studies published up to 1st December 2017 using terms relevant to the validation of EE estimates from activity monitors against criterion measures with the following strategy ((tracker AND EE) AND validation). The search was updated on 15th January 2018. The specific keywords and the full search strategy can be found in appendix 1.1. No language restrictions were applied and in the case of studies available only as an abstract, attempts were made to contact the authors to request the full text.

### 4.2.2 Inclusion and exclusion criteria

The analysis conducted here included only laboratory or field validation studies conducted in healthy adults (≥18 years) comparing a criterion measure of EE to an estimate of EE in kilocalories (kcal), kilojoules (kJ) or megajoules (MJ) from an activity monitor. Only wrist or arm-worn devices were included because there is a clear tendency towards wrist-worn devices amongst consumer devices and devices worn on alternative anatomical locations produce different accelerometry patterns and therefore estimates

of EE (Nelson et al., 2016). Studies must have reported concurrent EE estimates from one of the following criterion measures to be included: DLW, indirect calorimetry devices and metabolic chambers (Hills et al., 2014).

Adults with conditions deemed to produce atypical movement patterns were excluded, including Parkinson's disease, chronic obstructive pulmonary disease, cerebral palsy and amputees. These conditions are often associated with abnormal gait pattern and thus reduce accuracy in EE estimates (Van Remoortel et al., 2012). Devices requiring external sensors or components were also excluded. Studies reporting only accelerometer counts or studies involving post-hoc manipulation of the device output were excluded.

### 4.2.3 Study selection

Two authors (ROD and JT) independently assessed 100% of titles and abstracts for potential inclusion, with 10% screened independently by a third author (GF). In the case of disagreements between reviewers, the paper was retrieved in full-text and a mutual consensus was reached. Remaining articles were screened independently for inclusion at the full-text level by two authors (ROD and JT), with a third author (SS) screening 10%. Similarly, conflicts were resolved by discussion between reviewers.

### 4.2.4 Data extraction

From each of the included studies, characteristics of participants, validation protocol, criterion measure and the devices tested including model, wear site and output were extracted. Mean difference or EE estimates from the criterion measure and the device were extracted, along with standard deviation (SD), standard error (SE) or 95% confidence intervals (95% CI). If only SE was provided, SE was converted to SD. If data were not provided, authors were contacted to request the raw data. Where values were only presented in figures, a digitiser tool was used (Rohatgi, 2017). Data were extracted to a specialised spreadsheet and entered into Comprehensive Meta-analysis (CMA) (CMA, version 2; Biostat, Englewood, NJ) for analysis. All data extraction and data entry to CMA was performed by a single author (ROD) and was cross-checked for errors by a second author (JT).

### 4.2.5 Quality assessment

Risk of bias in included studies was determined using a modified version of the Downs and Black checklist for non-randomised studies (Downs & Black, 1998). The Downs and Black instrument is an established tool for determination of the quality of a study within a systematic review and meta-

analysis (Deeks et al., 2003). The modified version used in the present study carried a maximum score of 18 and was quantified as low (≤9, <50%), moderate (>9–14 points, 50–79%), or high (≥15 points, ≥80%) (MacDonald et al., 2016). The modified tool contained 17 questions, 10 related to reporting, three to external validity and four to internal validity. The risk of bias assessment was performed independently by two authors (ROD and JT), disagreements were resolved by discussion.

### 4.2.6 Statistical analysis

Descriptive statistics were calculated for studies included within the meta-analysis. The EE estimates from the device and criterion, SD or 95% CI, sample sizes and correlation coefficients for within-activity comparisons for each device were used to calculate effect sizes. Correlation coefficients were based on raw data from previously published studies or were conservatively estimated based on the mean of similar devices (See online supplementary material 3, (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020)). Where a study provided data for more than one comparison for one device, the selected outcomes were pooled to provide a single mean and prevent overpowering of a study. Hedges' g (ES) (Hedges, 1981) and 95% CIs were calculated using CMA, following the majority of studies in the literature testing the mean bias between activity monitors and criterion measures. A negative ES represents an underestimation relative to the criterion and a positive value represents an overestimation. Interpretation of ES was as follows: <0.20 as trivial, 0.20-0.39 as small, 0.40-0.80 as moderate and >0.80 as large (Cohen, 1977). A random-effects model was employed for all analyses based on the assumption that heterogeneity would exist between included studies due to the variability in study design (Higgins et al., 2009). To quantify heterogeneity, the $I^2$ statistic (Higgins & Thompson, 2002) was utilised and >75% was considered to represent large heterogeneity. To determine susceptibility to bias from one study, a leave one out analysis was conducted where the removal of one study would leave at least three studies. The study associated with the greatest change to the significance of the effect is reported. To assist interpretation of the error associated with each device, the percentage error was calculated using the percentage difference and weight within each meta-analysis.

### 4.2.7 Exploration of small study effects

To examine small study effects, data were visually inspected with funnel plots and subsequently, the effects were quantified by using Egger's linear

regression intercept (Egger et al., 1997). A significant Egger's statistic indicates the presence of a small study effect.

### 4.2.8 Moderators and subgroups

As well as overall, which represents a combination of all subgroups, subgroup meta-analyses were performed for specific activities/categories: 1) activity energy expenditure (AEE) which included comparisons of EE estimates from the device to a criterion during non-specific exercise protocols, circuits, arm ergometer, rowing and resistance exercises; 2) ambulation and stair climbing; 3) cycling; 4) running; 5) sedentary behaviours and household tasks and 6) total energy expenditure (TEE), representing comparisons to DLW. Moderator analyses were conducted between the sensors and all devices were grouped based on the inclusion of the following sensor hardware: 1) accelerometry alone (ACC); 2) heart rate alone (HR); 3) accelerometry and heart rate (ACC+HR); 4) accelerometry and heat-sensing or galvanic skin response (ACC+HS) and 5) accelerometry, heart rate sensors and heat-sensing or galvanic skin response sensors (ACC+HR+HS). Secondly, moderator analyses were conducted by the grade of devices. Devices produced by Actical, Actigraph and Bodymedia were considered as research-grade and all other devices included in the analysis were considered commercial devices. Comparisons between each moderator employed a random-effects model.

## 4.3 Results

A summary of the studies included in the systematic review is shown in appendix 1.2. Four studies could not be synthesised by meta-analysis as the mean difference between activity monitors and criterion measurements were not provided (Lopez et al., 2018; Machač et al., 2013; Reeve et al., 2014; Shcherbina et al., 2017). The remaining studies were included in the meta-analysis (Alsubheen et al., 2016; Bai et al., 2018; Benito et al., 2012; Berntsen et al., 2010; Berntsen et al., 2011; Bhammar et al., 2016; Boudreaux et al., 2018; Brazeau et al., 2016; Brazeau et al., 2011, 2014; Brugniaux et al., 2010; Calabro et al., 2015; Calabró et al., 2014; Casiraghi et al., 2013; Chowdhury et al., 2017; Colbert et al., 2011; Correa et al., 2016; Diaz, Krupka, Chang, Peacock, et al., 2016; Diaz, Krupka, Chang, Shaffer, et al., 2016; Dondzila & Garner, 2016; Dooley et al., 2017; Drenowatz & Eisenmann, 2011; Erdogan et al., 2010; Fruin & Rankin, 2004; Furlanetto et al., 2010; Gastin et al., 2018; Heiermann et al., 2011; Imboden et al., 2018; Jakicic et al., 2004; Johannsen et al., 2010; Kim & Welk, 2015; King et al.,

2004; Koehler et al., 2011b; Lee et al., 2011; Lee et al., 2014; MacKey et al., 2011; Martien et al., 2015; McMinn et al., 2012; Melanson et al., 2009; Montoye, Mitrzyk, et al., 2017; Murakami et al., 2016; Nelson et al., 2016; Papazoglou et al., 2006; Price et al., 2017; Reece et al., 2015; Rousset et al., 2015; Ryan & Gormley, 2013; Slinde et al., 2013; Smith et al., 2012; Soric et al., 2012; St-onge et al., 2007; Stackpool et al., 2014; Tucker et al., 2015; Van Hoye et al., 2014, 2015; Vanhelst et al., 2012; Vernillo et al., 2015; Wahl et al., 2017; Wallen et al., 2016; Woodman et al., 2017). A total of 1946 participants were included, with a mean age of 35 years (range 20 to 86 years). The mean BMI was 24.9 kg/m$^2$ (range 21.8 to 31.6 kg/m$^2$). Within the included studies, 104 comparisons between devices and a criterion were included. This represented 58 commercial and 46 research-grade device comparisons. ACC was comprised of 35 comparisons, 1 in HR devices, 20 in ACC+HR devices, 45 in ACC+HS and 3 in ACC+HR+HS. Concerning the activity performed, 35 comparisons were classed as AEE, ambulation and stairs included 55 comparisons, 23 were cycling tasks and 38 were running tasks. Sedentary and low-intensity was comprised of 30 comparisons and TEE included 16 comparisons.

### 4.3.1 Devices

A total of 40 devices were tested in the included studies. One device was forearm-worn, 6 were worn on the upper arm (triceps) and 33 were wrist-worn. Characteristics of the devices, the number of studies and weighted percentage error for each device is shown in appendix 1.2.

### 4.3.2 Meta-analysis

Individual study effect sizes and allocation to moderator variables are provided in the online supplement for this publication (See online supplementary material 6 (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020)). A minimum of three comparisons was required for meta-analysis and as such, pooled ES for individual devices or moderators where three or more comparisons were available are reported. Statistical outputs for each device are presented online (See online supplementary material 7 (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020)).

### 4.3.3 Quality assessment

The modified Downs and Black scores revealed a median score of 13, with one study being classed as low quality (Melanson et al., 2009), 48 classed as moderate and 11 classed as high quality (supplementary materials 8). The questions included in the modified tool and the percentage of studies

fulfilling each question is shown online (See online supplementary material 9, (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020)).

### 4.3.4 Overall

A forest plot of individual devices for all activities is shown in figure 4.2. Overall, devices underestimated EE (ES: −0.23,95% CI −0.44 to −0.03; n=104; p=0.03) and showed significant heterogeneity between devices ($I^2$ =92.18%; p=<0.001). Significant underestimations relative to criterion measures were observed for the Garmin Vivofit (GVF; ES: −1.09, 95% CI−1.60 to −0.57; n=5; p<0.001) and the Jawbone UP24 (ES:−1.16, 95% CI −1.78 to −0.54; n=3; p<0.001).). The SenseWear Armband Pro3 (SWA p3) also underestimated EE (ES:−0.32. 95% CI −0.62 to −0.01; n=12; p=0.04).). Sensitivity analysis revealed that the removal of six comparisons altered the significance of the SWA p3 (p>0.05), the most influential of which decreased the ES to -0.19 (95% CI: -0.50 to 0.11; p=0.21) (Soric et al., 2012). The Apple watch (AW) Bodymedia CORE armband (BMC), Fitbit charge HR (FCHR), Fitbit Flex (FF), Jawbone UP (JU), Nike Fuelband (NF), SenseWear Armband (SWA) SenseWear Armband Pro2 (SWA p2), and Mini (SWAM) did not differ significantly from criterion measures. However, sensitivity analysis showed the FCHR differed significantly with the removal of one study (ES: 0.34, 95% CI: 0.20 to 0.49; p<0.001) (Wallen et al., 2016). The NF was the only device that did not display significant heterogeneity between studies ($I^2$ =25.44%; p=0.26), with the remaining devices having $I^2$ values ≥ 66.91% (all p≤0.05). No device showed evidence of small-study effects.

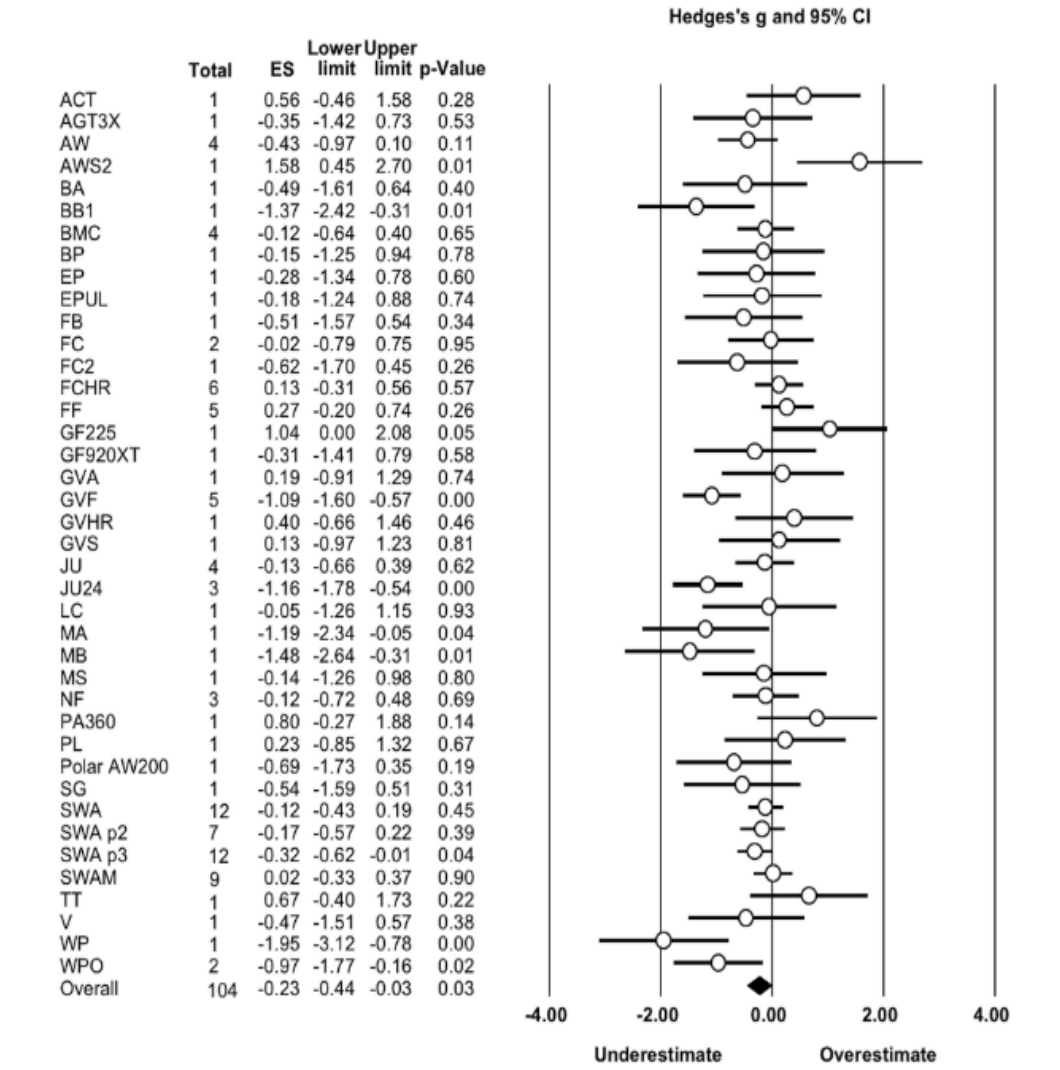



**Figure 4.2** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the overall comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.5 Activity energy expenditure

A forest plot of individual devices during activities classed as AEE is shown in figure 4.3. For AEE, the pooled estimate of all devices was a non-significant tendency to underestimate EE compared with criterion measures (ES: -0.34, 95% CI: -0.71 to 0.04; n=35; p=0.08) and significant heterogeneity was observed between devices ($I^2$ =94.96%; p<0.001). The SWA p2 underestimated EE (ES: -0.78, 95% CI: -1.48 to -0.08; n=3; p=0.03)

and had moderate, non-significant heterogeneity ($I^2$ =64.19%; p=0.06). The BMC, NF, SWA and SWAM did not differ significantly from criterion measures but all displayed significant heterogeneity. No device showed evidence of small-study effects.

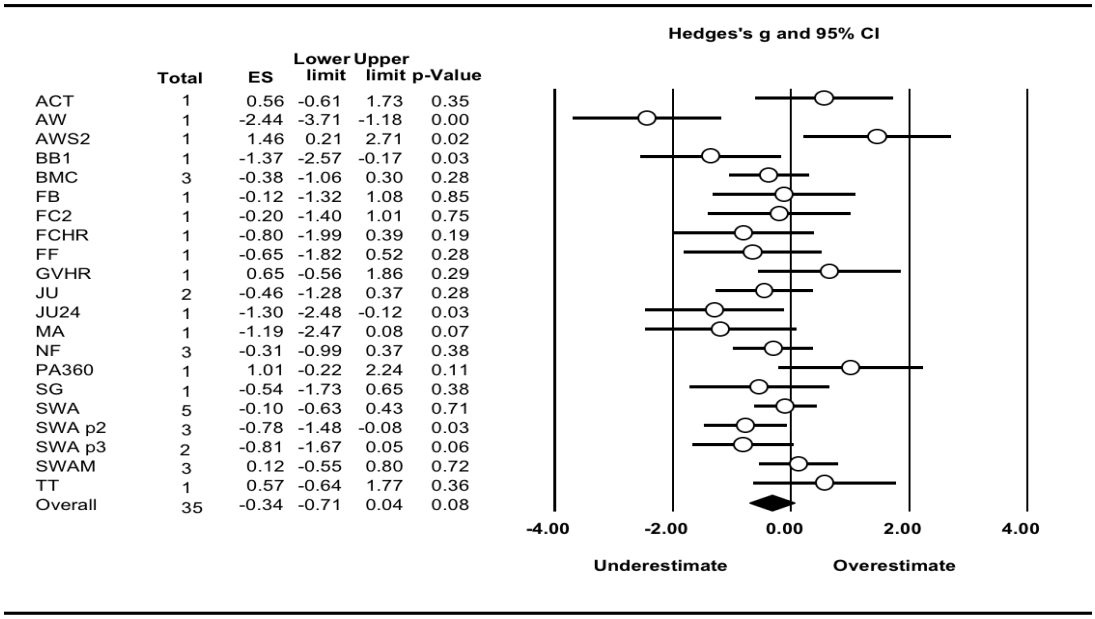



**Figure 4.3** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the activity energy expenditure comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.6 Ambulation and stairs

A forest plot of individual devices during ambulation and stair climbing is shown in figure 4.4. The pooled estimate of all devices did not differ from criterion measures (ES: -0.09, 95% CI: -0.45 to 0.27; n=55; p=0.62) and significant heterogeneity was observed between devices (I2 =93.74%; p<0.01). The FCHR (ES: 0.78, 95% CI 0.28 to 1.29; n=5; p=0.002) and FF (ES: 1.10, 95% CI: 0.43 to 1.77; n=3; p=0.001) overestimated EE. The GVF underestimated EE (ES: -1.24, 95% CI: -1.86 to -0.62; n=4; p<0.001), however, sensitivity analysis revealed that the removal of two comparisons significantly altered the mean effect (p>0.05) the most influential significantly altered the mean effect to ES: -1.32 (95% CI: -2.73 to 0.08; p=0.07) (Alsubheen et al., 2016). Further, there was evidence of small-study effects (intercept= -13.76, 95% CI: -19.72 to -7.80; p=0.01). The SWA

overestimated EE (ES: 0.79, 95% CI: 0.25 to 1.33; n=5; p=0.004) and sensitivity analysis revealed that the removal of four comparisons significantly altered the mean effect (p>0.05) the most influential significantly altered the mean effect to ES: 0.33 (95% CI: -0.26 to 0.92; p=0.28) (Gastin et al., 2018). The AW, JU, SWA p3 and SWAM did not differ significantly from criterion measures. The mean effect of the SWAM was significantly altered by the removal of two studies; the removal of the most influential study yielded a significant overestimation (ES: 0.57, 95% CI: 0.20 to 0.94; p=0.003) (Wahl et al., 2017). All devices showed significant heterogeneity.

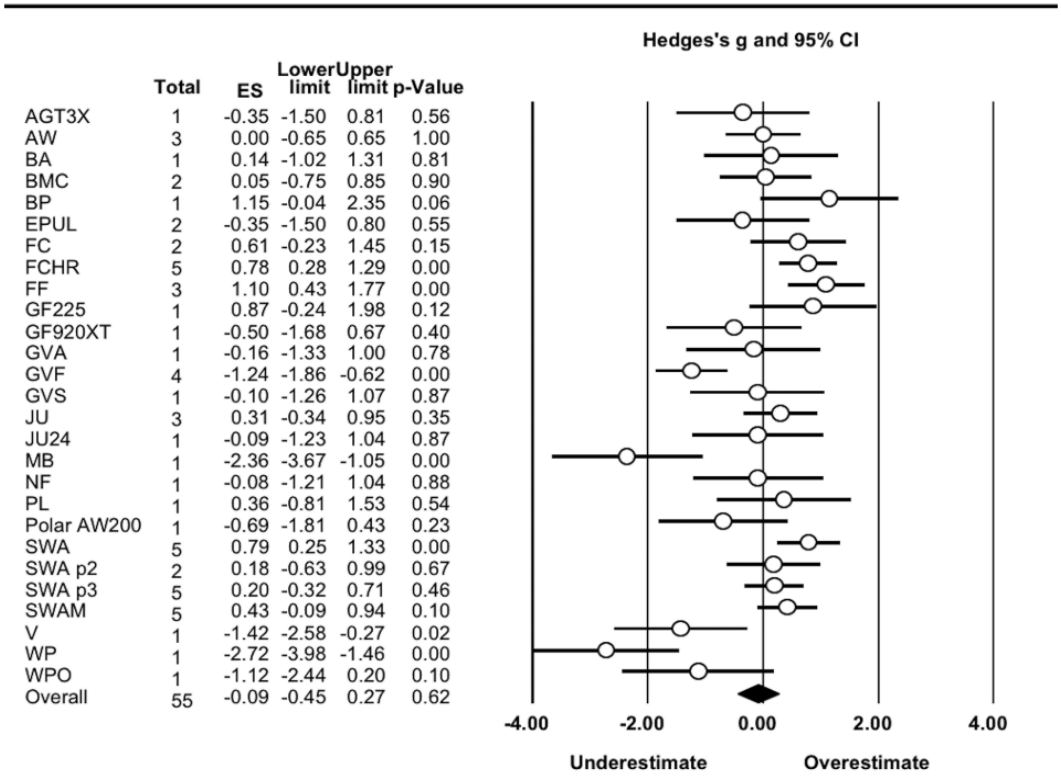


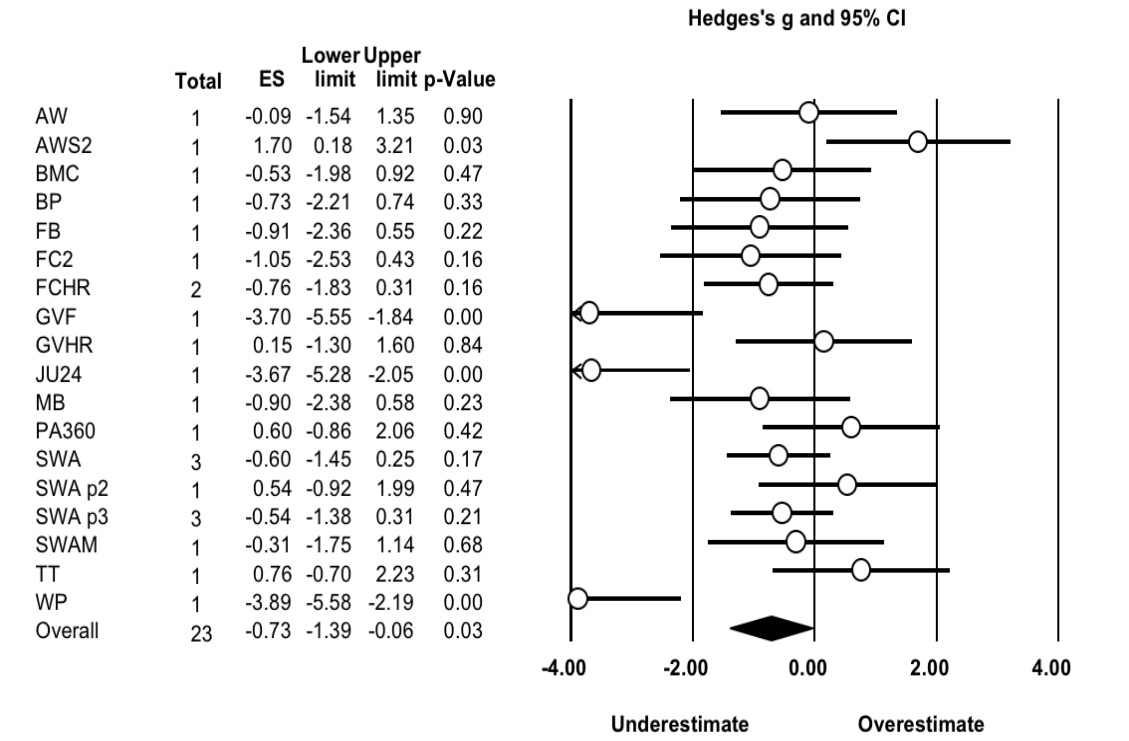

**Figure 4.4** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the ambulation and stairs comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.7 Cycling

A forest plot of individual devices during cycling is shown in Figure 4.5. The pooled estimate of all devices was significantly lower than criterion measures (ES: -0.73, 95% CI: -1.39 to -0.06; n=23; p=0.03) and significant

heterogeneity was observed between devices ($I^2$ =94.74%; p<0.01). The SWA did not differ significantly from criterion but showed significant heterogeneity ($I^2$ =89.39%; p<0.001). The SWA p3 did not differ from criterion measures and showed moderate heterogeneity ($I^2$ =54.95%; p=0.11).




**Meta Analysis Cycling**

**Figure 4.5** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the cycling comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.8 Running

A forest plot of individual devices during running is shown in Figure 4.6. The pooled estimate was not statistically different from criterion measures (ES: -0.08, 95% CI: -0.41 to 0.25; n=38; p=0.65) and significant heterogeneity was observed between devices (I2 =92.05%; p=<0.001). The FCHR, GVF and SWA did not differ from criterion measures. Sensitivity analysis revealed the

removal of one study changed the overall effect for the FCHR (ES: 0.59, 95% CI: 0.28 to 0.90; p<0.001) (Wahl et al., 2017). Significant heterogeneity was observed for the FCHR (I2 =66.8%; p=0.03) and SWA (I2 =96.79; p<0.001), but not for the GVF (I2 =46.39%; p=0.15).

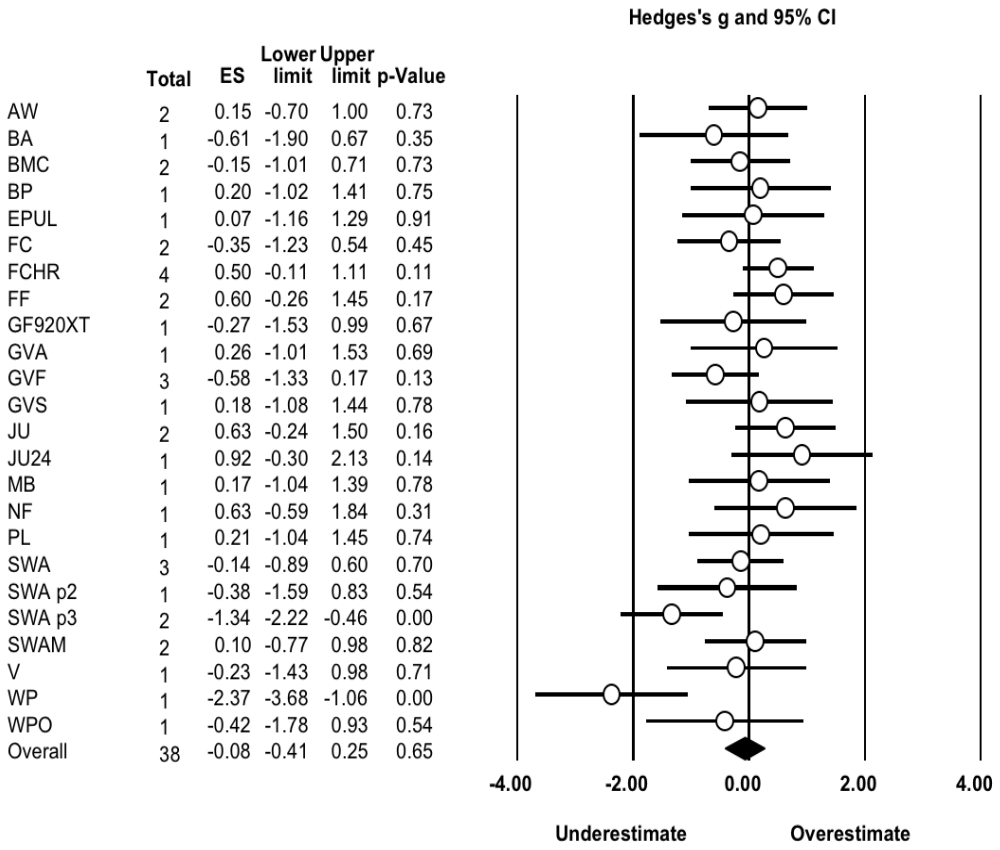



**Figure 4.6** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the running comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.9 Sedentary and household tasks

A forest plot of individual devices during sedentary and household tasks is shown in figure 4.7. The pooled effect was not statistically different from criterion measures (ES: -0.09, 95% CI: -0.51 to 0.32; n=30; p=0.66) and

significant heterogeneity was observed between devices (I2 =94.84%; p<0.001). The AW, FCHR and SWAM were not statistically different from criterion measures. The SWA p3 overestimated EE (ES: 0.67, 95% CI: 0.00 to 1.34; p=0.049). Sensitivity analysis revealed that the removal of three studies changed the mean effect, the most influential of which decreased the ES to 0.41 (95% CI: -0.01 to 0.82; p=0.05) (Brazeau et al., 2014). Observed heterogeneity was significant for the AW, SWA p3 and SWAM. The FCHR had moderate, non-significant heterogeneity (I2 =59.60%; p=0.06).

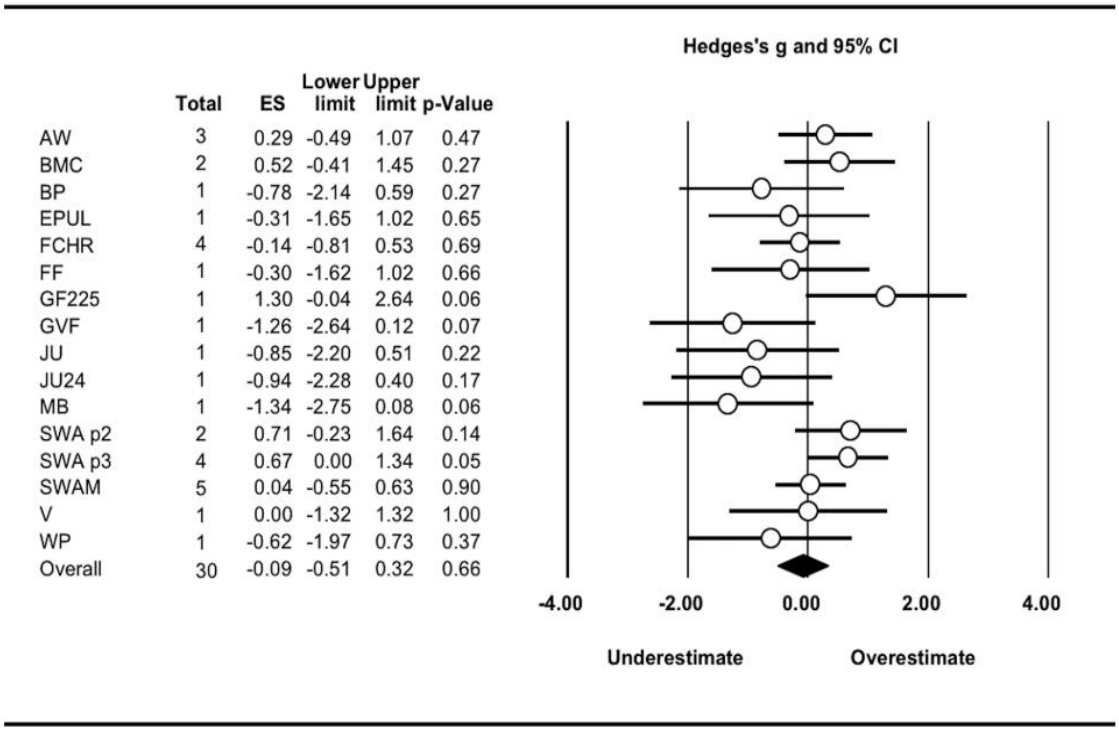



**Figure 4.7** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the sedentary and household comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.10 Total energy expenditure

A forest plot of individual devices for the measurement of TEE is shown in Figure 4.8. The pooled effect for TEE showed a significant underestimation of EE (ES: -0.68, 95% CI: -1.15 to -0.21; n=16; p=0.005) and significant heterogeneity was observed between devices ($I^2$ =92.17%; p<0.01). The SWA p3 did not differ significantly from criterion measures and showed significant heterogeneity ($I^2$ =94.20%; p=0.001).

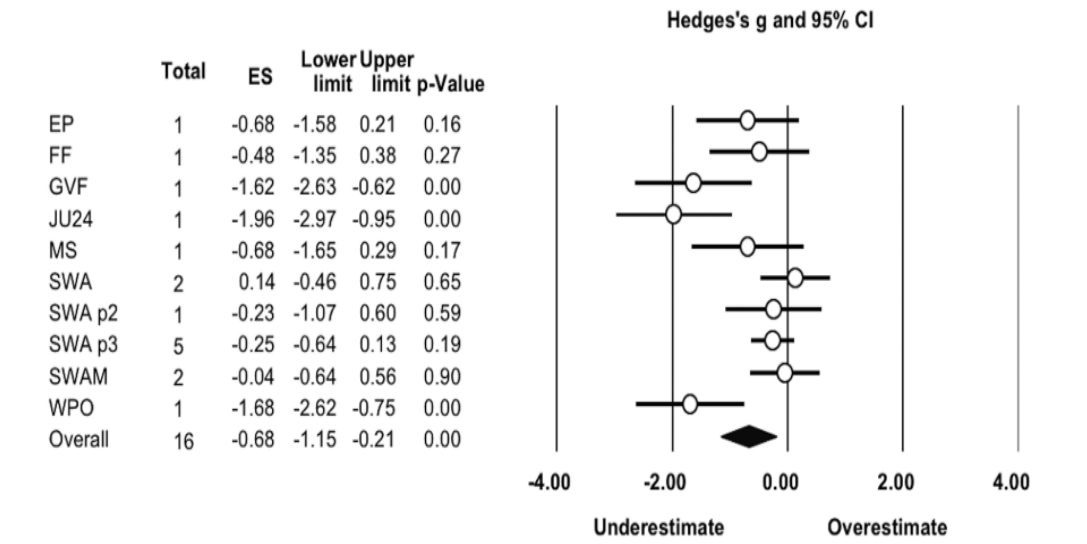



**Figure 4.8** Pooled Hedges' g and 95% confidence intervals (CI) for estimates of energy expenditure relative to the criterion for each device for the TEE comparison.

Total refers to the number of effect sizes. A negative Hedges' g statistic represents an underestimation and a positive Hedges' g represents an overestimation.

### 4.3.11 Moderator analyses

The results of moderator analyses are shown in table 4.1. Overall, there was a significant difference between sensors (p=0.003). The pooled estimate of EE from ACC+HR and ACC+HS was not statistically different from criterion but ACC+HS showed a non-significant tendency for underestimation, and ACC and ACC+HR+HS both significantly underestimated EE. In the AEE comparison, there was no statistical difference between sensors, but ACC+HS significantly underestimated EE, ACC showed a non-significant tendency for underestimation and ACC+HR did not differ significantly from criterion measures. During ambulation and stair climbing, a significant difference between sensors was observed, with estimates of EE from ACC+HR and ACC+HS being significantly higher than the criterion. In

cycling, significant differences were observed between sensors, with ACC devices underestimating EE. During running activities, none of the pooled mean estimates were significantly different from criterion. For sedentary and household tasks, a significant difference was observed between sensors; ACC+HR was not different from criterion measures whereas ACC and ACC+HS underestimated and overestimated EE respectively. For TEE, sensors differed significantly; ACC underestimated EE, whereas ACC+HS did not differ significantly from criterion.

When analysed by commercial and research-grade devices, no significant difference was observed overall, for AEE, cycling or running. For both the ambulation and stairs comparison and the sedentary and household tasks comparison, commercial devices were closer to criterion measurements, with research-grade devices significantly overestimating. For TEE, research-grade devices were superior, with commercial devices significantly underestimating EE.

**Table 4.1** Moderation analysis for the level of sensors and grade of the device by subgroup.

Data are shown where at least 3 comparisons were included. *P*-value refers to a between subgroup comparison. *Significant effect size at the subgroup level ($p$<.05).

| Moderator variable | Subgroup level | *p*-value | Hedges' g (95% CI) |
|---|---|---|---|
| **Overall activities** | | | |
| Sensors | ACC (n=35) | <0.01 | -0.36 (-0.55, -0.17)* |
| | ACC + HR (n=20) | | 0.06 (-0.18, 0.31) |
| | ACC + HR + HS (n=3) | | -0.99 (-1.65, -0.33)* |
| | ACC + HS (n=45) | | -0.15 (-0.32, 0.01) |
| Device grade | Commercial (n=58) | 0.27 | -0.27 (-0.42, -0.12)* |
| | Research (n=46) | | -0.14 (-0.31, 0.03) |
| **AEE** | | | |
| Sensors | ACC (n=8) | 0.19 | -0.40 (-0.84, 0.04) |
| | ACC + HR (n=9) | | -0.04 (-0.47, 0.38) |
| | ACC + HS (n=16) | | -0.32 (-0.63, -0.01)* |
| Device grade | Commercial (n=18) | 0.62 | -0.38 (-0.67, -0.08)* |
| | Research (n=17) | | -0.27 (-0.57, 0.04) |
| **Ambulation and stairs** | | | |
| Sensors | ACC (n=24) | 0.01 | -0.23 (-0.51, 0.06) |
| | ACC + HR (n=10) | | 0.44 (0.02, 0.87)* |
| | ACC + HS (n=19) | | 0.40 (0.08, 0.72)* |
| Device grade | Commercial (n=35) | 0.05 | -0.04 (-0.28, 0.20) |
| | Research (n=20) | | 0.37 (0.05, 0.68)* |
| **Cycling** | | | |
| Sensors | ACC (n=3) | <0.01 | -3.75 (-4.65, -2.85)* |
| | ACC + HR (n=9) | | -0.03 (-0.47, 0.40) |
| | ACC + HS (n=9) | | -0.41 (-0.84, 0.02) |
| Device grade | Commercial (n=14) | 0.28 | -0.82 (-1.30, -0.35)* |
| | Research (n=9) | | -0.41 (-0.99, 0.17) |
| **Running** | | | |
| Sensors | ACC (n=19) | 0.18 | -0.06 (-0.36, 0.24) |
| | ACC + HR (n=7) | | 0.34 (-0.15, 0.82) |
| | ACC + HS (n=10) | | -0.36 (-0.77, 0.05) |
| Device grade | Commercial (n=28) | 0.08 | 0.06 (-0.18, 0.30) |
| | Research (n=10) | | -0.36 (-0.76, 0.04) |
| **Sedentary and household** | | | |
| Sensors | ACC (n=6) | <0.01 | -0.56 (-1.16, -0.13)* |
| | ACC + HR (n=9) | | 0.14 (-0.27, 0.55) |

| | | | |
|---|---|---|---|
| | ACC + HS (n=13) | | 0.39 (0.06, 0.73)* |
| Device grade | Commercial (n=17) | <0.01 | -0.27 (-0.59, 0.05) |
| | Research (n=13) | | 0.41 (0.05, 0.77)* |
| **TEE (DLW)** | | | |
| Sensors | ACC (n=5) | <0.01 | -1.24 (-1.66, -0.81)* |
| | ACC + HS (n=10) | | -0.13 (-0.40, 0.14) |
| Device grade | Commercial (n=6) | <0.01 | -1.13 (-1.51, -0.76)* |
| | Research (n=10) | | -0.13 (-0.39, 0.14) |

## 4.4 Discussion

Given the clinical and consumer uptake of wrist and arm-worn activity monitors, which can be used for the estimation of EE, the aims of this meta-analysis were (i) to determine the relative accuracy of current devices for the estimation of EE when compared to criterion measures (aim 1, objective 1) (ii) to investigate the importance of specific sensors within devices (aim 1, objective 2) and (iii) to compare commercial and research-grade devices.

For devices with sufficient comparisons to be analysed separately from the main pooled effect, significant error relative to criterion measures was observed for Garmin, Fitbit, Jawbone and Bodymedia/SenseWear products. Garmin, Fitbit and Jawbone represent a major share of the commercial wearable market at the time of writing (Price et al., 2017) and iterations of Bodymedia/SenseWear products are widely used in research and have been since 2004 (Jakicic et al., 2004). Whilst it is initially encouraging that the ES for many devices was not significantly different from the respective criterion, the 95% CI observed in many cases indicates the potential for these devices to produce erroneous estimates of EE and as such none of these devices should be considered sufficiently accurate. The SenseWear armband Mini was the most accurate device overall but error reported in studies ranged from -21% to 15%, and the 95% CI spans from -0.33 to 0.37, indicating that the measure could be improved. Studies in this analysis followed the manufacturer's instructions for setup, with researchers ensuring the position of the device and characteristics such as height, weight, sex and age were correct. In free-living environments, the lack of researcher presence could yield greater error than observed in this analysis (Evenson et al., 2015), as indicated by the moderate, significant underestimation for the pooled effect in the TEE subgroup analysis.

An accurate yet affordable measure of free-living EE, with a measure of the change in energy storage, can be used to estimate free-living EI in large cohorts (Sanghvi et al., 2015). In this context, TEE may be considered the most important activity subgroup in this meta-analysis. However, as described in chapter 1, TEE is simply the sum of its components and the most variable and unpredictable component is EE during activity (Hills et al., 2014). In agreement with previous studies (Calabró et al., 2014; Dondzila & Garner, 2016; Woodman et al., 2017), the accuracy of devices differs by activity and this may be related to the inability of devices to differentiate between activity types. The Fitbit Charge HR was the most tested commercial device in this analysis, and it showed a trivial, non-significant ES overall and during sedentary tasks but a moderate to large and significant overestimation during ambulatory activity. Considering that ambulatory activity is central to public health guidelines worldwide (Pollard & Wagnild, 2017) an error in the estimation of the EE associated with ambulation may have implications for estimates of TEE in a range of populations.

The observed error for different activity types may be because current algorithms do not take physical activity type or bodily posture into account (Schneller et al., 2015). Indeed, activity recognition is considered an important direction for wearable technology (Wright et al., 2017) and has been used to improve estimates of EE (Welk et al., 2007). Montoye et al have shown that accelerometers worn on the wrists and thigh can be used to predict activity type (Montoye et al., 2016) and similarly, the SenseWear software employs complex pattern-recognition algorithms to determine activity type (Calabró et al., 2014). It may be that such activity recognising capabilities contribute to the trivial or small ES observed for the SenseWear Armband Mini in all comparisons. The challenges associated with activity recognition have been reviewed previously (Plasqui, 2017) and as this technology develops, activity-specific EE prediction equations may offer the opportunity to reduced errors associated with activity types.

### 4.3.1 Sensors

A 2012 review concluded that multisensory and triaxial accelerometry devices improve estimates of EE, relative to uniaxial devices (Van Remoortel et al., 2012). Due to recent technological advancements, triaxial accelerometry, as well as heart rate or heat sensing technology are now commonplace in wearables (Chowdhury et al., 2017) but it was unclear exactly whether the same pattern of improvement would emerge in the present analysis. It was hypothesised that the addition of this technology to

accelerometry would improve estimates of EE and the current results indicate that this is the case. It is established that accelerometry is limited for non-weight-bearing activities (Van Hoye et al., 2014), and again, this was confirmed by the observation that accelerometry underestimated EE during cycling activities. Significant underestimations were also observed during sedentary and household tasks and TEE, which is likely a product of the limited arm movements associated with some of these activities.

Accelerometer and heart rate devices moderately overestimated EE during ambulation and stair climbing. Some of this error may be attributable to the individual variability in the relationship between heart rate and EE. Individual calibration of this relationship in the Actiheart device is associated with improved estimates of EE (Brage et al., 2007) and may offer a means for further reducing the error observed in wrist and arm-worn devices. An alternative explanation for this is the variability in estimates of heart rate from photoplethysmography heart rate sensors. A recent study reported a small mean error of -5.9 bpm in the Fitbit Charge 2, but wide limits of agreement of -28.5 to 16.8 bpm (Benedetto et al., 2018) and this variability is a common finding (Bai et al., 2018; Boudreaux et al., 2018).

### 4.3.2 Device Grade

The third aim of this meta-analysis was to compare commercial and research-grade devices. Commercial devices may be developed with affordability and comfort as a primary focus, and as a consequence, it may be unreasonable to expect commercial devices to match the validity of research-grade devices. Recent consumer monitors share similar technology with established research-grade multi-sensor devices (Chowdhury et al., 2017) and the positive implications this can have for EE estimates can be seen in the present results. A benefit of research-grade devices for TEE was observed, but commercial devices were statistically superior in ambulation and during sedentary tasks, thus this hypothesis is not completely supported. The present results question the use of wrist or arm-worn research-grade devices for the validation of newer devices. Comparisons to criterion measures such as DLW or indirect calorimetry are more appropriate when absolute accuracy is required (Hills et al., 2014). Further, it is important to highlight that other research-grade devices, for instance, the Actiheart, which is worn on the chest (Brage et al., 2007), are likely to be more accurate than research-grade devices included in this study (Chowdhury et al., 2017). Further research is needed to establish whether

research-grade devices that are worn in other locations such as the chest, hip or thigh outperform consumer-based devices.

### 4.4.3 Limitations

Separate pooled analyses to determine the accuracy of individual activity monitors were performed for a limited number of devices due to the small number of comparisons available for the remaining devices (i.e., less than three comparisons). This limitation is inevitable considering the large number of activity monitors included in this review. Nevertheless, the inclusion of all devices in the overall pooled analysis provides an extensive and robust evaluation of the difference in EE outcomes between activity monitors and criterion measures.

The majority of analyses conducted within this review demonstrated large heterogeneity within and between devices. Such heterogeneity is not unexpected and in many cases may be attributable to a disparity in the protocols employed (Higgins, 2008). The rate of EE is likely to be elevated in the period following higher intensity exercise and the inclusion of only the steady-state period may influence the extent to which devices differ from criterion measures (Gastin et al., 2018). There is also the possibility that the discrepancy between device estimates relates to populations studied (Koehler & Drenowatz, 2017). As few devices currently provide open-access EE algorithms, the potential for this to create heterogeneity remains uncertain. Despite the heterogeneity, the statistically significant outcomes in many cases suggest a consistent direction in effect sizes.

External validity was low in 46 studies pooled in this meta-analysis, which must be considered when interpreting the present results. It must also be noted that the present analysis was limited to healthy individuals and therefore our results cannot be generalised to populations with conditions that produce abnormal gait patterns. Lastly, there is a lag between product release and testing in research environments (Boudreaux et al., 2018) and some of the devices included in this meta-analysis are no longer in production so the continued validation of newer devices is imperative.

## 4.5 Conclusion

This meta-analysis collated studies evaluating the validity of EE estimates by wrist or arm-worn devices. Devices vary in accuracy depending on activity type and the significant heterogeneity means caution must be exercised when interpreting these results. Devices with heart rate sensors often

produced better estimates than devices using accelerometry only. However, this was not consistent across all activities. Wrist and arm-worn research-grade devices were more accurate than commercial devices for estimates of TEE but researchers should be aware that such devices do not guarantee superior accuracy. Future research should aim to understand and reduce the error in EE estimates from wrist or arm-worn devices in different activity types. This may be achieved through activity recognition techniques, incorporating physiological measures and exploring the potential for individual calibration of these relationships.

# Chapter 5 – A validation study of the Fitbit charge 2 for the measurement of energy expenditure and heart rate

## 5.1 Introduction

The case has been made throughout this thesis that whilst gold standard measures for EE are available, they are associated with constraints which prevent their use at scale. Unlike indirect calorimetry or expensive stable isotopic measures, wearable devices can provide estimates of EE over small epochs, in ecologically valid environments and large populations. This could bring a new dimension to the assessment of free-living EE across a range of population groups in health and disease. Despite this promise, inaccurate instruments are undesirable as they may bias the outcomes and conclusions of a study (Dhurandhar et al., 2015). Indeed, the previous chapter confirmed that the accuracy of almost all currently available devices is variable between activity types, and this is consistent with an existing body of literature validating wearable devices (Dooley et al., 2017; Drenowatz & Eisenmann, 2011; Evenson et al., 2015; Feehan et al., 2018).

The release of new commercial devices is often faster than validation studies (Boudreaux et al., 2018) and thus, the accuracy of newer devices remains unclear. Physiological sensors, including heart rate sensors (Yang & Hsu, 2010) are commonplace in newer activity monitors (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020) and such innovation may be bringing the accuracy of commercial devices in line with more established research-grade devices (Chowdhury et al., 2017). The reason heart rate monitoring is important to the assessment of EE is that a linear relationship exists between oxygen consumption ($VO_2$) and heart rate during moderate to high-intensity activities (Ceesay et al., 1989; Spurr et al., 1988) and therefore monitoring heart rate at the minute-level enables relative physical activity intensity estimates (Karvonen et al., 1957; Schrack et al., 2018) or EE (Achten & Jeukendrup, 2003) to be estimated. As such, it is not unexpected to observe that multivariate approaches, in which physiological and movement variables are incorporated into predictive algorithms, improve the estimation of physical activity or EE relative to accelerometry alone (Brage et al., 2015; O'Driscoll, Turicchi, Beaulieu, et al., 2020). Of course, the integrity of EE estimates depends on the validity of the heart rate estimates in the populations and activities of interest.

### 5.1.1 Chapter aims

The purpose of the present study was to evaluate the heart rate and EE estimates of the FB, a modern commercial-grade wearable device. The second aim was to validate the EE estimates of the research-grade SWA. Validations are conducted during sedentary, household, ambulatory and cycling tasks in a heterogeneous population.

## 5.2 Methods

### 5.2.1 Participants

A total of 59 participants were enrolled in this study (age range: 22-73 years, weight range 49.2 - 105.99 kg) and participant characteristics are presented in table 5.1. Participants were primarily recruited from the Leeds centre of the NoHoW trial (n = 44; see **section 3.1.3**). A further 15 participants were recruited from the local area. For further details on the inclusion/exclusion criteria and ethics, please see **sections 3.1.1, 3.2.1** and **3.3.1**

**Table 5.1** Descriptive characteristics of the included sample.

FM = Fat mass, FFM = Fat free mass, RMR = Resting metabolic rate, SBP = Systolic blood pressure, DBP = diastolic blood pressure. Data are shown as means ± SD.

| | N | Age | Weight | FM% | FFM% | FM (kg) | FFM (kg) | RMR (kcal) | SBP | DBP | Resting heart rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 59 | 44.41 ± 14.1 | 75.7 ± 13.6 | 32.5± 10.3 | 67.5 ± 10.3 | 24.8 ± 10.7 | 49.8 ± 8.9 | 1581.8 ± 280.4 | 121.9 ± 11.5 | 78.1 ± 8.5 | 64.9 ± 10 |
| **F** | 41 | 46.6 ±13.1 | 71.5 ± 12.9 | 35.6 ± 8.8 | 64.4 ± 8.8 | 26.3 ±10.6 | 45.2 ±4.8 | 1466.6 ± 223.9 | 118.4 ± 11.5 | 77.4 ± 9 | 67.1 ±10 |
| **M** | 18 | 39.8 ±15.5 | 84.7 ±10.6 | 24.5 ± 9.9 | 75.5 ± 9.9 | 21.1 ± 10.5 | 61.7 ± 4.7 | 1830.5 ± 225.6 | 129.4 ± 7 | 79.5 ± 7.5 | 60.3 ± 8.6 |

### 5.2.2 Protocol

Following body composition and RMR measurements, participants transitioned to the exercise laboratory where a physical activity protocol was performed. Participants were initially seated for 5 minutes, followed by 5 minutes standing. Next, participants performed 5 minutes of treadmill walking, incline walking (4 km/h, 5% incline), running and incline running (6-8 km/h, 5% incline). Participants were then given a 3-minute resting period and then transitioned to a cycle ergometer and performed 5 minutes of low-intensity (30 watts), and moderate-intensity cycling (60 watts). Lastly, after another resting period, participants performed a 5-minute folding task and a 5-minute sweeping task. Throughout this protocol, participants wore a polar heart rate monitor, FB and a SWA at all times whilst breath by breath respiratory data was collected using a stationary metabolic cart (see below).

### 5.2.3 Physical measurements

Participants arrived at the laboratory in the morning and in a fasted state having refrained from the intake of food, caffeine and exercise in the 12 hours prior. After completing a medical screening questionnaire and providing informed consent, height was measured without shoes using a stadiometer (Leicester height measure, SECA; UK). Blood pressure and resting heart rate were measured using an automatic sphygmomanometer (Microlife BP A2 Basic, Gentle Technology, Microlife, Clearwater, FL, USA, Inc.). Next, body composition was estimated using a 2-compartment model via air displacement plethysmography (BodPod, Life Measurement, Inc.; USA), as described in **section 3.4.1.5**. The RMR of each subject was obtained by the method described in **section 3.4.2.2**. If RMR data were unavailable (n=2), RMR was estimated with a body mass index specific RMR algorithm (Müller et al., 2004).

### 5.2.4 Wearable monitors

Several wearable devices were used in this study and more detailed descriptions of them can be found in **chapter 3**. Heart rate was assessed during the physical activity protocol using a Polar m400 heart rate monitor watch (Polar Electro, Kempele, Finland) and a Polar H7 chest strap (Polar Electro, Kempele, Finland), which transmitted second-level data via a Bluetooth connection (See **section 3.4.3.4**). The FB (Fitbit Inc, San Francisco, CA, USA), a wrist-worn activity monitor was worn on the non-dominant wrist in this study to estimate EE and heart rate (see **section 3.4.3.1**). Lastly, the SWA (BodyMedia Inc., Pittsburgh, PA) was fitted by an

elastic strap around the non-dominant arm and was used to estimate EE rate (See **section 3.4.3.2**).

### 5.2.5 Vyntus CPX (Jaeger)

A stationary metabolic cart fitted with a respiratory facemask (Vyntus CPX, Jaeger-CareFusion, UK) was used as the criterion measure of EE in the present study. Full details of the experimental methodology for this system can be found in **section 3.4.2.3.**

### 5.2.6 Statistical analysis

All analyses were conducted in R version 3.5.1 and Rstudio Version 1.1.447. Statistical significance was accepted at $p < 0.05$ for all analyses. Descriptive statistics (mean ± SD) were calculated for age, weight, height, FM, FFM and RMR. Data from the devices and criterion measures were averaged to provide a mean heart rate in beats per minute (BPM) or EE (kcal/min$^{-1}$) for each participant. Data for each of the outputs were matched by time for each participant. Next, the first minute of data from each activity performed in the activity protocol was removed leaving minutes 2-5, which were considered steady-state. These data were then averaged for each participant's activity bout and this figure was used in analyses. Analyses for each of the devices, heart rate and EE were conducted separately. In line with previous research (Bai et al., 2018) a range of statistical tests were used. Firstly, the agreement between the criterion measure and devices was assessed with Pearson's correlation coefficient. Second, the method of Bland-Altman (Altman & Bland, 1983) was used to investigate the mean difference between criterion and device estimates, using the 'BlandAltmanLeh' package in R. Root mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) (See **section 3.5.4.1**), were calculated. Lastly, 'equivalence tests' were conducted. These methods are explained in further detail in **section 3.5.4.1**. Differences in absolute percentage error were investigated by analysis of variance (ANOVA) and a post-hoc Tukey honest significant difference test, with homogeneity of variance assessed with Levene's test, from the 'car' package in R. The relationship between continuous variables (age, RMR, height, weight, FM, FFM, resting heart rate, systolic and diastolic blood pressure) and absolute error rate in EE and heart rate was investigated with Pearson's correlations, using the 'cor' function, from the 'stats' package in R.

## 5.3 Results

The physical activity protocol was performed by all participants (n = 59) however the running task (n=49), the 5% incline run (n=30) and the moderate cycling tasks (n=58) were not performed by all participants due to ranges in physical fitness within the sample.

### 5.3.1 Fitbit Charge 2

Synchronisation errors occurred for two participant's FB data and therefore 57 participants data were included in FB analyses. The pooled result of all available bouts showed a mean overestimation in EE by the FB of 0.8 (kcal/min$^{-1}$), RMSE = 2.3 (kcal/min$^{-1}$), correlation coefficient of r = 0.77, MAPE = 44% and a non-significant equivalence test ($p > 0.05$) indicating that the FB was not equivalent to the criterion measure overall. The activity-specific statistics and the number of bouts included in the analyses are presented in table 5.2. The poorest accuracy was observed in the folding and sweeping tasks, in which the FB overestimated EE. The MAPE values were 93% and 81% for sweeping and folding, respectively (figure 5.1). The best accuracy and statistical equivalence was observed in incline running tasks (MAPE = 12%). A Bland-Altman plot of the overall error is shown in figure 5.2 (right), for which the 95% Limits of agreement were: -3.52, 5.14 (kcal/min$^{-1}$).

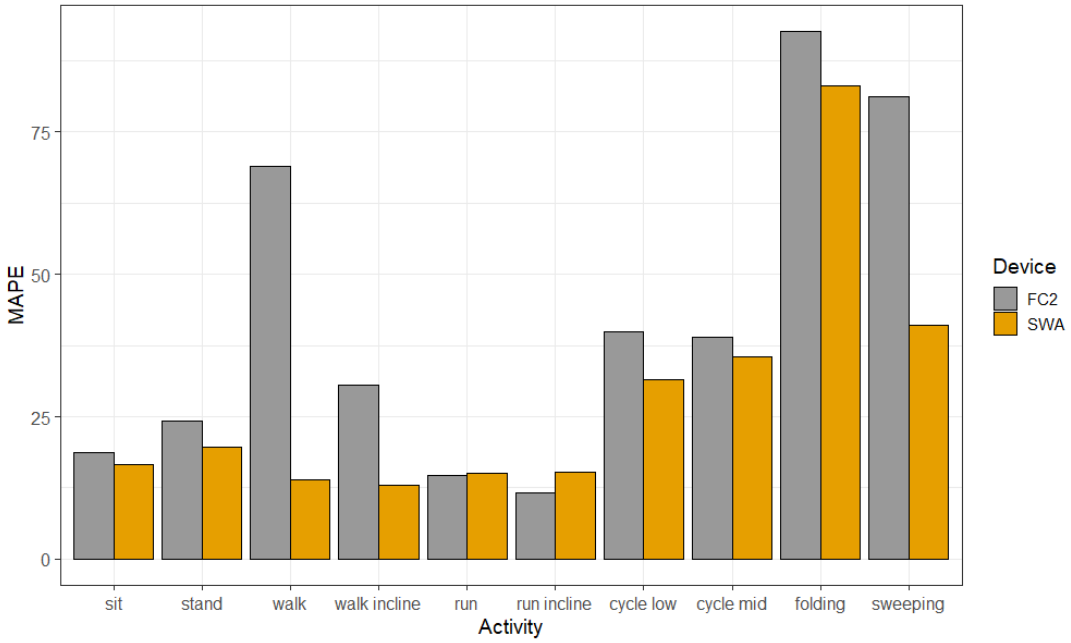



**Figure 5.1** A bar plot detailing the mean absolute percentage error (MAPE) of EE from the SWA (yellow) and FB (grey) for each of the activities performed in this study.

**Table 5.2** Statistics detailing the validity of EE estimates obtained from the FB and SWA.

Activity is laid out in the order dictated by the physical activity protocol. 'Bouts' refers to the number of minute activity bouts included, and 'ID' refers to the number of participants included in each comparison. 'Correlation' refers to Pearson's R. 'Equivalence' refers to the results of the equivalence tests and the absence of text implies a non-significant equivalence test. Data are shown as means ± SD. MAPE = Mean absolute percentage error, RMSE = Root mean squared error, MAE = Mean absolute error.

| Device | Activity | Bouts | ID | Device | Criterion | RMSE | MAPE | MAE | Correlation | Equivalence |
|---|---|---|---|---|---|---|---|---|---|---|
| **FC2** | sit | 228 | 57 | 1.08 ± 0.24 | 1.30 ± 0.31 | 0.32 | 19 | 0.26 | 0.66 | |
| | stand | 228 | 57 | 1.15 ± 0.29 | 1.47 ± 0.35 | 0.44 | 24 | 0.37 | 0.56 | |
| | walk | 228 | 57 | 7.10 ± 1.97 | 4.27 ± 0.86 | 3.35 | 69 | 2.83 | 0.39 | |
| | walk incline | 228 | 57 | 7.32 ± 2.39 | 5.66 ± 1.02 | 2.56 | 31 | 1.75 | 0.59 | |
| | run | 191 | 48 | 9.91 ± 1.91 | 9.18 ± 1.83 | 1.61 | 15 | 1.29 | 0.7 | |
| | run incline | 120 | 30 | 10.61 ± 2.57 | 11.14 ± 2.22 | 1.58 | 12 | 1.27 | 0.81 | Equivalent |
| | cycle low | 225 | 57 | 3.78 ± 2.17 | 4.49 ± 1.23 | 2.15 | 40 | 1.7 | 0.38 | |
| | cycle mid | 217 | 56 | 4.35 ± 2.50 | 5.59 ± 1.54 | 2.69 | 39 | 2.14 | 0.37 | |
| | folding | 228 | 57 | 5.57 ± 1.88 | 2.96 ± 0.61 | 3.11 | 93 | 2.7 | 0.42 | |

| Device | Activity | Bouts | ID | Device | Criterion | RMSE | MAPE | MAE | Correlation | Equivalence |
|---|---|---|---|---|---|---|---|---|---|---|
| | sweeping | 228 | 57 | 5.98 ± 1.69 | 3.38 ± 0.83 | 2.94 | 81 | 2.64 | 0.58 | |
| **SWA** | sit | 236 | 59 | 1.43 ± 0.31 | 1.29 ± 0.31 | 0.25 | 17 | 0.2 | 0.75 | |
| | stand | 236 | 59 | 1.67 ± 0.36 | 1.47 ± 0.34 | 0.33 | 20 | 0.26 | 0.71 | |
| | walk | 236 | 59 | 4.47 ± 0.79 | 4.28 ± 0.85 | 0.73 | 14 | 0.59 | 0.62 | Equivalent |
| | walk incline | 236 | 59 | 5.12 ± 0.82 | 5.67 ± 1.00 | 1.02 | 13 | 0.78 | 0.56 | |
| | run | 195 | 49 | 9.73 ± 1.99 | 9.18 ± 1.81 | 1.6 | 15 | 1.34 | 0.69 | |
| | run incline | 120 | 30 | 9.69 ± 1.94 | 11.14 ± 2.22 | 2.14 | 15 | 1.76 | 0.71 | |
| | cycle low | 233 | 59 | 3.17 ± 1.19 | 4.51 ± 1.22 | 1.63 | 31 | 1.4 | 0.7 | |
| | cycle mid | 225 | 58 | 4.13 ± 1.98 | 5.60 ± 1.52 | 2.42 | 35 | 1.93 | 0.41 | |
| | folding | 236 | 59 | 5.31 ± 2.18 | 2.97 ± 0.60 | 3.06 | 83 | 2.43 | 0.43 | |
| | sweeping | 236 | 59 | 4.33 ± 1.70 | 3.37 ± 0.82 | 1.8 | 41 | 1.3 | 0.43 | |

### 5.3.2 SenseWear Armband

All 59 participants data were available and were included in the SWA analyses. The pooled result of all available bouts was a mean overestimation of 0.03 (kcal/min$^{-1}$), RMSE = 1.7 (kcal/min$^{-1}$) correlation coefficient of r = 0.82, MAPE = 29% and a significant equivalence test ($p < 0.001$), indicating that the SWA was equivalent to the criterion measure overall. The activity-specific statistics and the number of bouts included in the analyses are presented in table 5.2. The SWA demonstrated the poorest accuracy in the folding task, in which it overestimated EE (MAPE = 83%). The lowest MAPE values were observed in the walking (MAPE = 14%) and walk 5% incline tasks (MAPE = 13%), which were overestimations and underestimations relative to the criterion measure, respectively. Equivalence testing showed statistical equivalence between the SWA and the criterion measure during walking only. A Bland-Altman plot of the overall error is shown in figure 5.2 (left), for which the 95% Limits of agreement were: -3.33, 3.38 (kcal/min$^{-1}$).

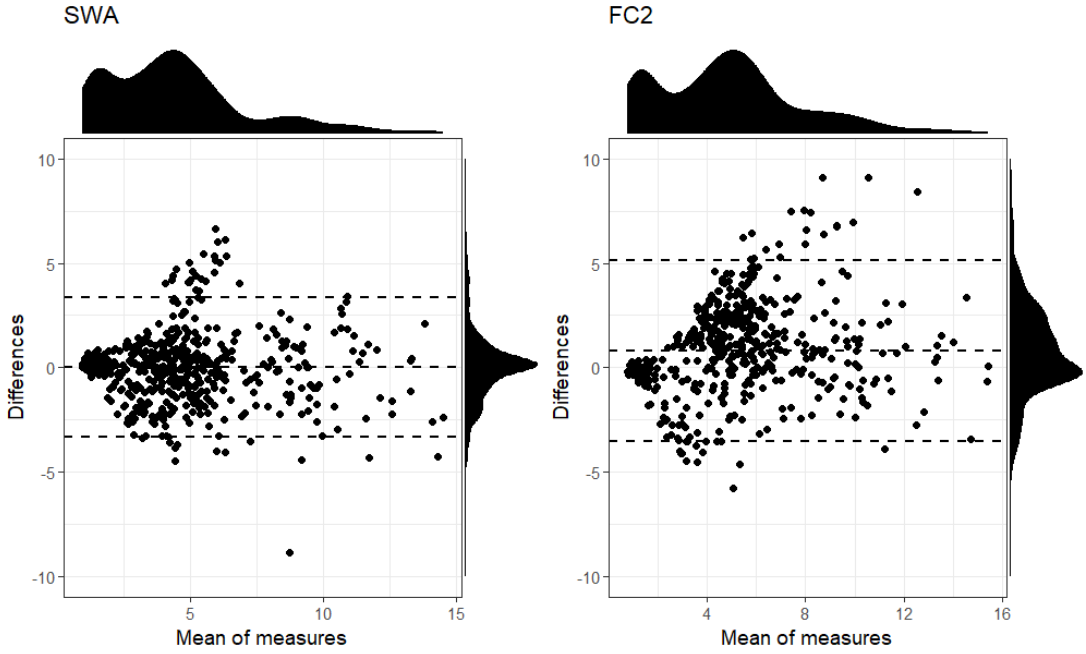



**Figure 5.2** Overall Bland-Altman plots of EE estimates from the SWA (left) and FB (right) relative to the criterion indirect calorimetry measure (Vyntus CPX).

Data are displayed as kcal/min. 'Differences' represents the difference between device and criterion estimates and is shown by the middle dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of

the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures.

### 5.3.3 Fitbit charge 2 heart rate

Polar heart rate connectivity error occurred for one participant and thus heart rate analyses were conducted with 56 of the 57 participants with FB data. The pooled result of all available bouts was 98 ± 27 BPM (polar) vs 99 ± 29 BPM (FB), RMSE = 20 BPM, correlation coefficient of r = 0.75, MAPE = 13% and a significant equivalence test ($p < 0.001$), indicating statistical equivalence. A Bland-Altman plot for errors in heart rate illustrates the agreement between criterion heart rate and FB heart rate by displaying the mean difference and 95% limits of agreement (figure 5.3), the 95% Limits of Agreement were: -37.94, 39.73 (BPM). Activity specific Bland-Altman plots are presented for all tasks in figure 5.4 and accuracy statistics are presented in table 5.3.

**Table 5.3** Statistics detailing the validity of heart rate estimates obtained from the FB, measured in beats per minute.

Activity is laid out in the order dictated by the physical activity protocol. 'Bouts' refers to the number of minute activity bouts included, and 'ID' refers to the number of participants included in each comparison. 'Correlation' refers to Pearson's R. 'Equivalence' refers to the results of the equivalence tests and the absence of text implies a non-significant equivalence test. Data are shown as means ± SD. MAPE = Mean absolute percentage error, RMSE = Root mean squared error, MAE = Mean absolute error.

| | Bouts (ID) | Device | Criterion | RMSE | MAPE | MAE | Correlation | Equivalence |
|---|---|---|---|---|---|---|---|---|
| **Sit** | 224 (56) | 62.29 ± 8.38 | 64.80 ±10.25 | 4.52 | 4 | 2.79 | 0.94 | Equivalent |
| **Stand** | 224 (56) | 66.44 ± 9.49 | 69.54 ±11.54 | 5.51 | 4 | 3.31 | 0.92 | Equivalent |
| **Walk** | 224 (56) | 101.80 ± 20.59 | 84.40 ±12.95 | 27.63 | 25 | 19.50 | 0.23 | |
| **Walk incline** | 224 (56) | 108.06 ± 22.94 | 97.19 ± 14.84 | 25.68 | 17 | 16.10 | 0.29 | |
| **Run** | 191 (48) | 136.15 ± 19.12 | 131.04 ± 20.93 | 17.16 | 8 | 10.02 | 0.66 | Equivalent |
| **Run incline** | 120 (30) | 142.13 ± 19.00 | 142.26 ± 20.15 | 11.85 | 5 | 6.81 | 0.81 | Equivalent |
| **Cycle low** | 217 (55) | 95.09 ± 20.55 | 105.40 ± 17.40 | 20.80 | 12 | 13.12 | 0.55 | |

| | Bouts (ID) | Device | Criterion | RMSE | MAPE | MAE | Correlation | Equivalence |
|---|---|---|---|---|---|---|---|---|
| **Cycle mid** | 209 (54) | 97.29 ± 24.44 | 114.73 ± 19.68 | 26.25 | 16 | 18.17 | 0.62 | |
| **Folding** | 224 (56) | 106.67 ± 12.91 | 102.03 ± 17.94 | 19.20 | 15 | 14.38 | 0.29 | Equivalent |
| **Sweeping** | 224 (56) | 102.27 ± 14.76 | 98.55 ± 18.78 | 20.17 | 16 | 14.40 | 0.31 | Equivalent |

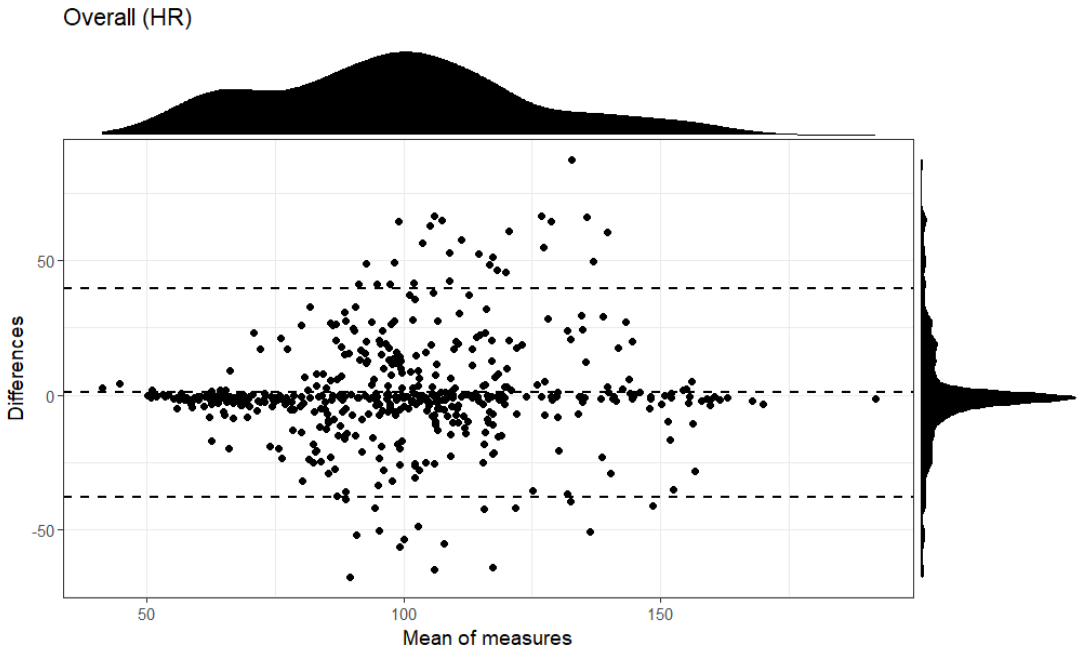



**Figure 5.3** Overall Bland-Altman plots of heart rate estimates from the FB relative to the criterion measure (Polar chest strap).

Data are displayed as beats per minute. 'Differences' represents device estimates – criterion estimates and is shown by the middle dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures

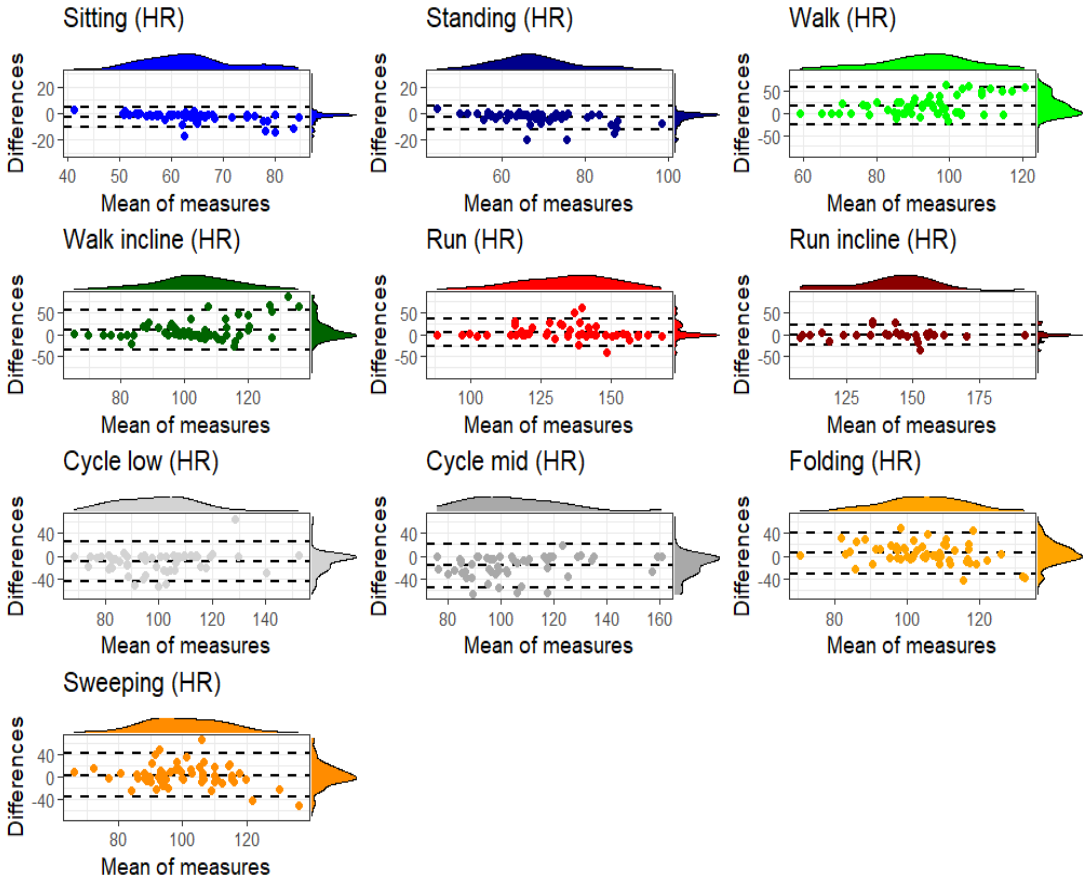



**Figure 5.4** Activity specific Bland-Altman plots for heart rate estimates from the FB relative to the criterion measure (Polar chest strap).

Activity specific Bland-Altman plots for heart rate estimates from the FB relative to the criterion measure (Polar chest strap). Data are displayed as beats per minute. 'Differences' represents device estimates – criterion estimates and is shown by the middle-dashed line. The upper and lower dashed lines represent the upper and lower 95% limits. Mean of measures represents the average value of the criterion and device estimate. The density plots visualise the distribution of data points over the differences between the measures and the means of the measures.

### 5.3.4 Predictors of absolute percentage error

Using the available data, no significant correlations were observed for any continuous variables and the absolute percentage error for heart rate and EE. ANOVA tests for the sex differences were not significant for absolute percentage errors in EE for the SWA or FB. In the heart rate comparison, a significant difference was observed between male bouts and female bouts with the absolute percentage error for males being significantly higher ($F$ = 4.158, p = 0.042).

## 5.4 Discussion

Commercial activity monitors can potentially be used to study the EE of free-living subjects, however, a critical barrier to their use is uncertainty regarding their accuracy. This study was necessary because scientific validations of the newest commercial devices (i.e. FB) are rare, especially in the context of more accurate devices such as the SWA. The study reported in this chapter investigated the validity of EE and heart rate estimates from the FB and EE estimates from the SWA. Comparisons for heart rate were made relative to a chest strap (Polar) and EE was compared to a stationary metabolic cart (Vyntus CPX). The principal findings are i) the research-grade SWA was observed to be more accurate than the commercial-grade FB overall ii) the heart rate estimates of the FB are generally in closer agreement with the criterion measures compared to EE estimates.

### 5.4.1 Energy expenditure

The FB, one of the newest Fitbit activity monitors, has been investigated previously for its validity in estimating EE, relative to indirect calorimetry (Boudreaux et al., 2018; Reddy et al., 2018), but this study provides a direct comparison with the SWA, a more established and commonly used, research-grade device. The results in this chapter substantiate previous research concluding that the SWA is more valid for the estimation of EE when compared to commercial activity monitors (Chowdhury et al., 2017; O'Driscoll, Turicchi, Beaulieu, et al., 2020). This being said, the SWA was not accurate across the range of activities performed, with MAPE values >25% in some activities. Specifically, low and moderate-intensity cycling, folding and sweeping.

Large overestimations were observed for the FB during household tasks. This most likely originates from the reliance on wrist accelerometry and this is a recognised limitation of devices located at this wear site (Ellis et al., 2016). Movements such as folding and sweeping involve rapid movements of the hand but are not particularly energetically demanding (typically ~4 METs) (Ainsworth et al., 2011). Importantly, the 'sedentary and household' meta-analysis reported in chapter 4 showed a non-significant effect for the Fitbit Charge HR (prior model to the FB device tested here). In this analysis, the FB underestimated sedentary EE but tends to overestimate household tasks. It could be that the effects are 'counter-balancing' each other in the meta-analysis, where the two activities were combined. The overestimation of EE in household tasks is opposite to the issue faced by more traditional

devices, which were worn on the hip and tended to underestimate the energy cost of tasks with limited ambulation (i.e. household tasks) (Hendelman et al., 2000; Nelson et al., 2016).

Notably, the MAPE values for the FB were lowest in running activities (indicating a high degree of accuracy) and higher during walking activities. This finding is reflective of the results of the analysis reported in **chapter 4**, in which the pooled results from five comparisons for the Fitbit Charge HR showed significant, moderate to large overestimation relative to criterion measures of EE for ambulatory activity and a non-significant overestimation during running (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020). It is not possible to confidently comment on the underlying cause of this error due to the proprietary nature of the algorithms. However, it is interesting to note that the greatest overestimate in heart rate estimates was also observed in the walking tasks. If heart rate is incorporated in the FB EE prediction algorithm, this could partially explain this result.

The performance of the SWA for the estimation of TDEE is well recognised (Casiraghi et al., 2013; Johannsen et al., 2010; Slinde et al., 2013). However, its accuracy in specific activity types is less established (Koehler & Drenowatz, 2017). Indeed, significant underestimations relative to indirect calorimetry when running at higher speeds (> 9.9 km/h) have been reported (Drenowatz & Eisenmann, 2011) and in a validation study involving cycling, the SWA again significantly underestimated EE (Koehler et al., 2011).The complementary results in comparisons to DLW may be largely influenced by the accuracy of the resting EE equations selected by the manufacturers, which are derived from participant characteristics (Nelson et al., 2016) and are generally specific to the population on which they were derived (Schofield et al., 2019). The present results offer some support for this supposition.

### 5.4.2 Heart rate

The conclusion that the estimates of heart rate from the FB are typically more accurate than EE estimates is reflective of previous research (Shcherbina et al., 2017; Wallen et al., 2016). When heart rate estimates were aggregated across all available bouts, the heart rate estimates of the FB were statistically equivalent to the criterion measure. Error in specific activity types was greater but the FB was statistically equivalent in most activity types. A recent study reported that erratic movements and a greater heart rate were associated with an increased error in heart rate (Nelson & Allen, 2019) and another concluded that the error was exacerbated with

increasing exercise intensity (Thomson et al., 2019). In contrast, the present results showed the highest error in the walking task, yet the greatest accuracy in the running tasks. The observation of the greatest error in walking is similar to that reported in a previous study investigating the Fitbit Surge device, where a greater error was observed during ambulatory tasks (Shcherbina et al., 2017). In contrast, two other studies investigating the FB report small underestimations in heart rate during walking (Nelson & Allen, 2019; Reddy et al., 2018).

No significant continuous correlates of the error for each device were identified and this includes body composition, which appears to be a novel investigation within this field. However, the percentage error in heart rate was significantly greater in males, when compared to females. Whilst the proprietary nature of the smoothing algorithms makes understanding the observed error challenging, photoplethysmography technology is likely to be influenced by device position and skin conditions which may differ between males and females (Stahl et al., 2016). Before the exercise condition, the position and tightness of the FB were standardised for all participants and it, therefore, seems unlikely that the position of the device played a role in the observed error. It remains to be seen whether the free-living performance of the FB will differ between participants in less controlled environments and this should be addressed in future research.

The seeming inability of the ‘out of the box’ FB estimates to accurately estimate EE is a primary limitation for energy balance research, particularly when the numerous benefits of cost, cloud storage and acceptance from participants are considered (Gualtieri et al., 2016; Wright et al., 2017). The results presented here indicate that it may be more appropriate to use commercial activity trackers, in their current format, to infer physical activity from step counts, or to estimate heart rate, as both of these metrics are typically observed to be more valid than EE estimates (Feehan et al., 2018). Alternatively, the application of metrics such as the heart rate reserve (Schrack et al., 2018), which can be used to define minute level relative intensity from heart rate data may be of greater use to researchers.

Considering that it is possible to access minute-level data from commercial wearables in many instances, this raises the possibility of the application of non-linear modelling to improve estimates of EE from commercial wearable devices. Advanced statistical learning techniques are being used to estimate EE and physical activity of tasks with better accuracy than linear regression approaches (Ellis et al., 2014; Montoye, Conger, et al., 2017; Staudenmayer

et al., 2009) and future research should investigate whether data from commercial activity monitors can be used to more accurately predict EE from sensor outputs. Further, the incorporation of body composition and participant characteristics into non-linear models could improve estimates of EE beyond the estimates of current activity monitors (Weyer et al., 1999).

### 5.4.3 Limitations

In this study, several different FB devices were used and data were synced with each participant's mobile phone application. The lack of standardisation of devices may be considered a limitation, as different firmware could potentially have been employed for different participants. However, this reflects the use of wearable devices in research environments, in which firmware updates are released sporadically. Secondly, whilst this study offers an analysis of the accuracy of two activity monitors for a relatively limited number of prescribed activities, it provides little insight into the ecological validity of these devices. Substantial over and underestimations from the FB, depending on the specific activity in question, were observed and therefore the error in free-living individuals will vary depending on the activities performed. Given that wearable devices will be used in free-living research, validation studies in free-living conditions are urgently required. Thirdly, this study was conducted in healthy, ambulatory individuals who were not pregnant, using medications associated with alteration to metabolic rate, and did not have cardiovascular, metabolic or renal disorders, illness or injury. The results may vary as the characteristics of study populations differ, however, except for gender difference in heart rate error, no evidence was found indicating that this was the case.

## 5.5 Conclusion

The SWA is more valid for the estimation of EE when compared to the commercial-grade FB, yet neither activity monitor can consistently estimate EE with equivalence to a criterion measure. The FB provides better estimates of heart rate than it does EE. The heart rate estimates are broadly, but not always, equivalent to criterion estimates across a range of activity types. It may therefore be more appropriate to focus on heart rate metrics for the assessment of physical activity, rather than EE in the FB. Mathematical models to estimate EI from bodyweight have been developed and validated (Sanghvi et al., 2015), and discussed extensively in **chapter 1** and **3**. However, these models make assumptions about the PAEE levels, which are unlikely to be constant between and within individuals during weight loss

and maintenance interventions (Kerns et al., 2017). An inexpensive, objective estimate of PAEE will therefore improve EI estimates from mathematical models and whilst devices such as the FB show large inaccuracies, it is likely that in their current form, they would be superior to an estimation of constant PAEE. This being said, the encouraging results in the heart rate analyses raise the possibility of estimating EE through new algorithms, which take the FB outputs (heart rate, movement etc) as input variables (Staudenmayer et al., 2009). Such approaches developed in an academic setting would be transparent concerning their development and assumptions.

# Chapter 6 – A methodology to account for missingness in physical activity data collected from commercial activity monitors

## 6.1 Introduction

The introduction of this thesis highlighted how technological advances in terms of size, data aggregation capabilities and the associated fall in cost has facilitated the use of tri-axial accelerometers in most new devices (Hills et al., 2014), creating opportunities for energy balance and related fields. The preceding chapters of this thesis have exclusively considered the errors associated with the sensor outputs or the derived EE estimates. However, accuracy is closely related to the amount of data available for analysis and this chapter seeks to investigate methods to minimise the biases brought about by missing data.

Missing data is a well-recognised phenomenon in accelerometer research (Troiano et al., 2008) and it is attributable to behavioural (i.e. removal for aesthetic reasons) and non-behavioural factors (i.e. device technical failures, charging). Non-wear time in accelerometery research has previously been detected by defining periods in which the signal of acceleration in each axis falls below a threshold for some time, often a predefined period between 10-120 minutes (Choi et al., 2011; Ridgers & Fairclough, 2011). Researchers then permit a maximum amount of non-wear time per day, which may be up to 14 hours (Tudor-Locke et al., 2012). The aim of defining such a period is to determine the amount of missing data which minimally influences the inferences of the study (Liu et al., 2016). It is also common to define a minimum number of valid days within a measurement period and if these criteria are met, an average or total value for physical activity metrics can be estimated (Doherty et al., 2017; Kapteyn et al., 2018).

Missing accelerometer data may detrimentally influence the conclusions of a study in several ways. If EE or physical activity is calculated from incomplete data, true physical activity or EE may be under-estimated (depending on the assumptions made about missing data). If missing periods occur in individuals that differ behaviourally or demographically from those with more complete data then the study's conclusions may be compromised (Loprinzi et al., 2013). A range of strategies have been developed to limit the bias introduced by missing accelerometer data (Stephens et al., 2018). These

methods make use of the observed (non-missing) data to build predictive models of missing data points and have utilised mean imputation (Meng et al., 2020), combined multivariate strategies (Lee, 2013; Staudenmayer et al., 2012) or normalisation by the amount of wear-time (Chen et al., 2009; Katapally & Muhajarine, 2014).

Some differences exist between research and commercial-grade devices which prevent the application of previous strategies to devices such as the FB; First, commercial activity monitors are cloud-connected, facilitating the assessment of physical activity for longer periods than research-grade equivalents, which typically measure physical activity maximally over a single week (Thraen-Borowski et al., 2017), meaning that subjects need to remove the devices for recharging. Commercial activity monitors are also increasingly equipped with heart rate monitoring devices (Benedetto et al., 2018), which can facilitate the estimation of the relative intensity of physical activity or EE, through heart rate reserve (HRR) or flex methodologies (Bassett et al., 2012; Rennie et al., 2001; Schrack et al., 2018; Silva et al., 2015) but also creates different patterns of missingness. For example, missing data may be identified through the loss of contact with the wrist (and therefore no measured heart rate), indicating that the device has most likely been removed. This results in the detection of smaller windows of removal, compared to longer periods used when the accelerometer signal is the determinant of missingness (Choi et al., 2011; Ridgers & Fairclough, 2011). These differences highlight an important need to develop methods to limit the bias associated with missing data from these devices. There has been no attempt to develop or apply imputation methodologies to commercially available multisensory activity monitors (e.g. FB).

### 6.1.1 Chapter aims

The purpose of the present study was to propose and evaluate a methodology designed to minimise the bias introduced by missing data collected from a commercial activity monitor (FB). Firstly, a series of intra-class correlation analyses were performed to investigate the minimum data required to achieve a reasonably non-biased aggregation of physical activity data collected by a FB. Next, the results of an autocorrelation analysis are presented, which serve as the rationale for the development of a method which scales temporally proximate data to produce summaries over a given measurement period. Lastly, in a series of simulation experiments using real datasets with simulated missingness, the performance of the proposed methodology was compared to alternative imputation strategies.

## 6.2 Methods

### 6.2.1 Participants

Data were collected as part of the NoHoW trial (ISRCTN88405328) (Scott et al., 2019) (Detailed in **sections 3.1.1, 3.1.2** and **3.1.3**). For the simulation experiments conducted in this study, FB data from 109 participants each wearing a FB for 14 days (minutes = 2,197,440, hours = 36,624, days = 1526) were used. This sample was selected based on the quantity of non-wear time (<2.5% data missing within the first 14 days). Utilising a sample with minimal degrees of missingness allows 'true', near-complete data to be held back for comparison with imputation methods.

### 6.2.2 Fitbit Charge 2 (FB)

All participants enrolled in the NoHoW trial were provided with a FB (Fitbit Inc, San Francisco, CA, USA), which is detailed in **section 3.4.3.1**. In the present study, non-wear time is defined by the absence of a heart rate measure and all devices were set to 'auto' mode by default, which ensured that no heart rate reading was transmitted when the device was not on the wrist.

### 6.2.3 Autocorrelation analyses

The algorithm proposed in this study was initially based on a series of autocorrelation analyses which are presented below. In autocorrelation analyses, the correlations between values in the time series are computed as a function of the time lag between them, defined in minutes in this case. For these analyses, autocorrelation values for time lags of up to 7 days (10080 minutes) were calculated for each participant individually, indicating time points within a week with the highest correlation. Figure 6.1 illustrates the autocorrelation for steps and heart rate for 90 minutes and 10081 minutes, respectively.

The average of the autocorrelation values (ACF) reached within 60 minutes for steps were: 15 mins: ACF = 0.31, 30 mins: ACF = 0.21, 45 mins: ACF = 0.15, 60 mins: ACF = 0.12, comparatively, heart rate values are higher: 15 mins: ACF = 0.62, 30 mins: ACF = 0.52, 45 mins: ACF = 0.46, 60 mins: ACF = 0.41. Although there is evidence of periodic patterns on subsequent days, the value does not exceed ACF = 0.09 for steps, which is observed at a lag of 1441 minutes and ACF = 0.25 is observed for heart rate at 1440 minutes,

the differences in these values are likely attributable to the stochastic nature of steps when compared to heart rate. Notably, the value at 10081 mins (7 days) is ACF = 0.05 for steps and ACF = 0.13 for heart rate. Thus, the greatest autocorrelation values are observed locally for both steps and heart rate.

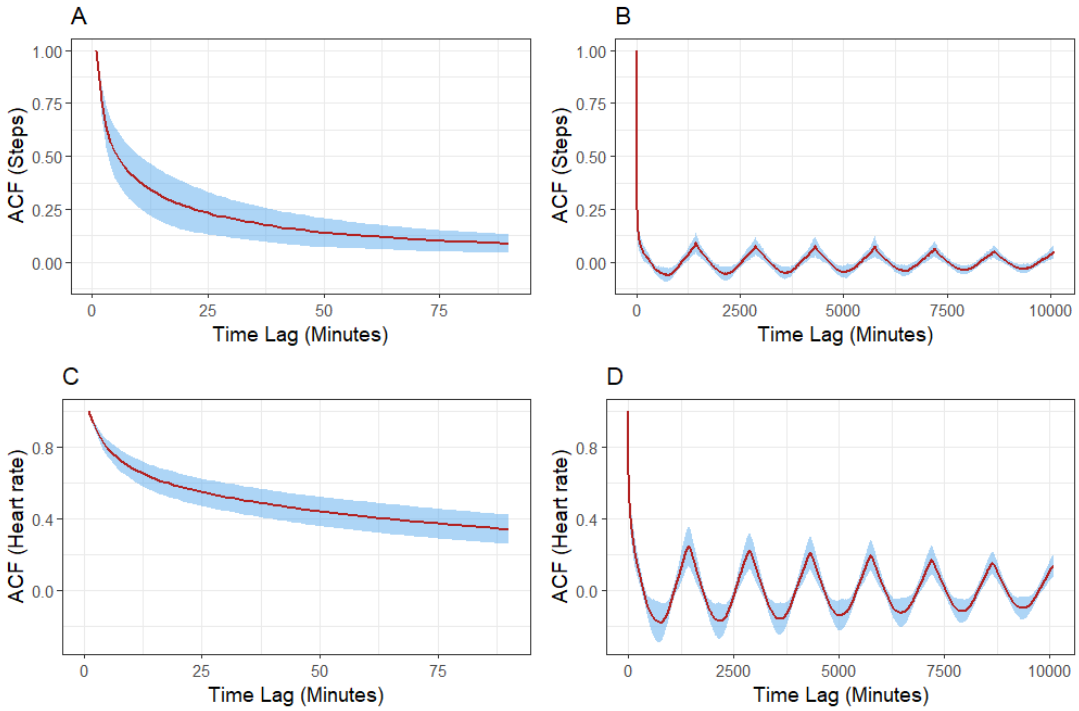



**Figure 6.1** Autocorrelation (ACF) values for steps with time lags of 90 minutes (A), 10,080 minutes (B) and heart rate with time lags of 90 minutes (C) and 10,080 minutes (D).

Average ACF values are shown in red and the blue ribbon represents ± 1 standard deviation.

### 6.2.4 Wear time requirements

To investigate the minimum amount of wear-time required for a valid hour, day or 14-day period, intraclass correlation (ICC) analyses were conducted, as ICC is a widely used and accepted means of determining measurement agreement (Koo & Li, 2016). In each of these experiments, data were deleted incrementally and at random and the ICC was calculated between the partially deleted data and the 'true' steps at each increment. An ICC threshold of 0.9 was used as the selection criterion, to align with a previous related publication in the Biobank study (Doherty et al., 2017). The first investigation was to determine the minimum time required within a single hour with adjustment for wear time, and thus the remaining data was divided by the proportion of the wear time and this adjusted value was used for ICC

analyses. In the daily and 14-day analyses, adjustments for wear time were not made. For all analyses, two-way mixed-effects agreement models were used (Koo & Li, 2016) and this was conducted with the 'icc' function from the 'rel' package in R. Figure 6.2a demonstrates that if 5 minutes of data are present and scaled to 60 minutes, the ICC threshold of 0.9 is reached. In the daily analysis, the ICC threshold was met at 18-19 hours per day (Figure 6.2b). It is important to note that the ICC comparisons for each day include non-scaled data despite using scaled data in the algorithm (outlined below). When scaling by the proportion of wear time per day, the number of hours required will be lower. 18 hours were used to ensure that true data are available from different parts of the day (i.e. morning, afternoon, evening) and this is a conservative requirement in line with previous research (Shook et al., 2018). To establish minimum 14-day requirements, the ICC threshold was met at 3 days (Figure 6.2c). For the final algorithm, 4 days were required including at least one weekend day as the minimum criteria for inclusion, owing to the potential for differential patterns of physical activity between weekdays and weekend days (Shiroma et al., 2019).

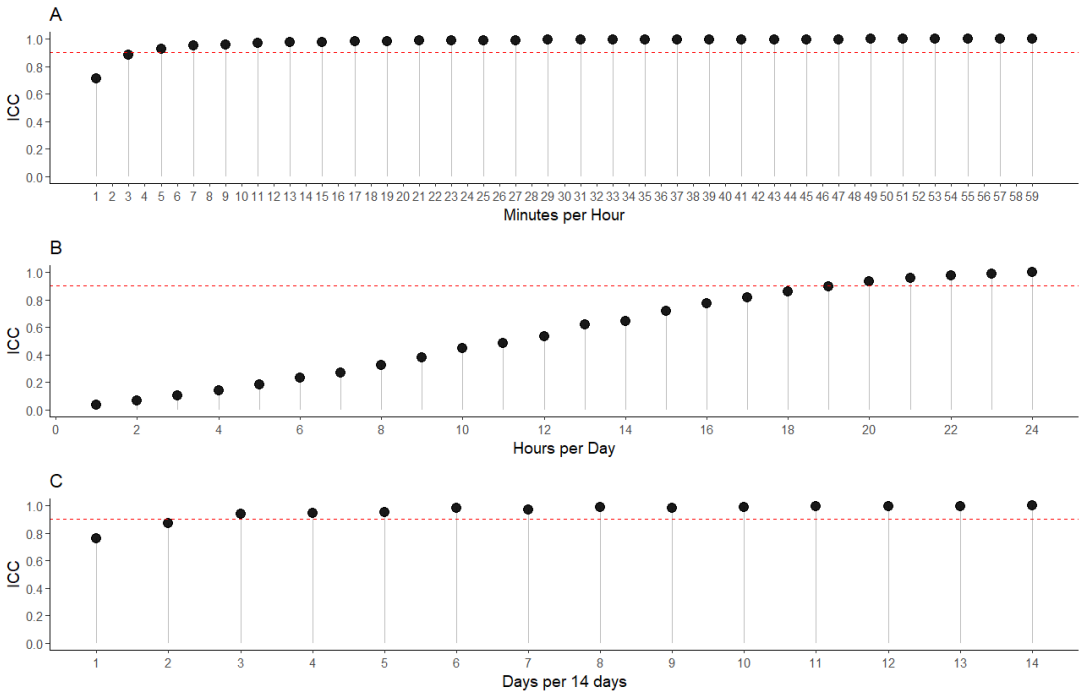



**Figure 6.2** Intraclass correlations (ICC) for incrementally deleted data and 'true' data. Data are presented for scaled minutes per hour (A), for hours per day (B) and for the number of days per 14 days (C).

### 6.2.5 NoHoW algorithm

Based on these analyses a scaling algorithm is proposed, referred to from hereon as ‘NoHoW algorithm’ as follows:

1: *If* non-missing minutes per hour < 5 *then* remove hour from dataset *else* sum available minutes to provide hourly total

2: Divide the number of available minutes per hour by 60 to give the proportion of wear time per hour

3: Divide hourly total by the proportion of wear time per hour to provide a scaled hourly total

4: *If* available hours per day < 18 *then* remove day from dataset *else* sum all available hours to give daily total

5: Divide the number of available hours by 24 to give proportion of wear time per day

6: Divide daily total by the proportion of wear time per day to provide a scaled daily total

7: *If* available days per 14 days < 4 *or* < 1 weekend day *then* remove 14-day period from dataset *else* average all valid days

### 6.2.6 Simulation study 1

In order to test the algorithm, two simulation experiments were performed. In the first experiment, traditional imputation methods and the proposed algorithm were tested. This was achieved by creating datasets with simulated missingness from each of the included participant’s true data and holding back this true data to be compared to the imputed datasets. The time point at which the data were removed was random and the length of each deleted period was uniformly sampled between one and 120 minutes in duration. The decision to insert missing data at random positions was informed by observing the proportion of missing FB data for each hour in the first 14 days of the NoHoW study, on average 22.83% was missing with a range of 21.1% at 13:00-13:59 to 25.96% at 23:00-23:59 (Figure 6.3). To determine the length of missing periods in this study, the length of each missing period in the first 14 days of the NoHoW study was determined, where the length was less than an entire day (1440 minutes). Of the 146,165 missing periods,139,213 (95.24%) were less than 60 minutes and 3882 (2.7%) were greater than 120 minutes (Figure 6.4), thus 120 minutes was

set as the upper limit for the length of insertions. The final parameter in the missing data algorithm was the number of missing periods, which was set to 40. This resulted in the amount of missing data per day being 13.7% (11.76% inserted) on average and ranging up to 44.4% (36.81% inserted) in simulation study 1.

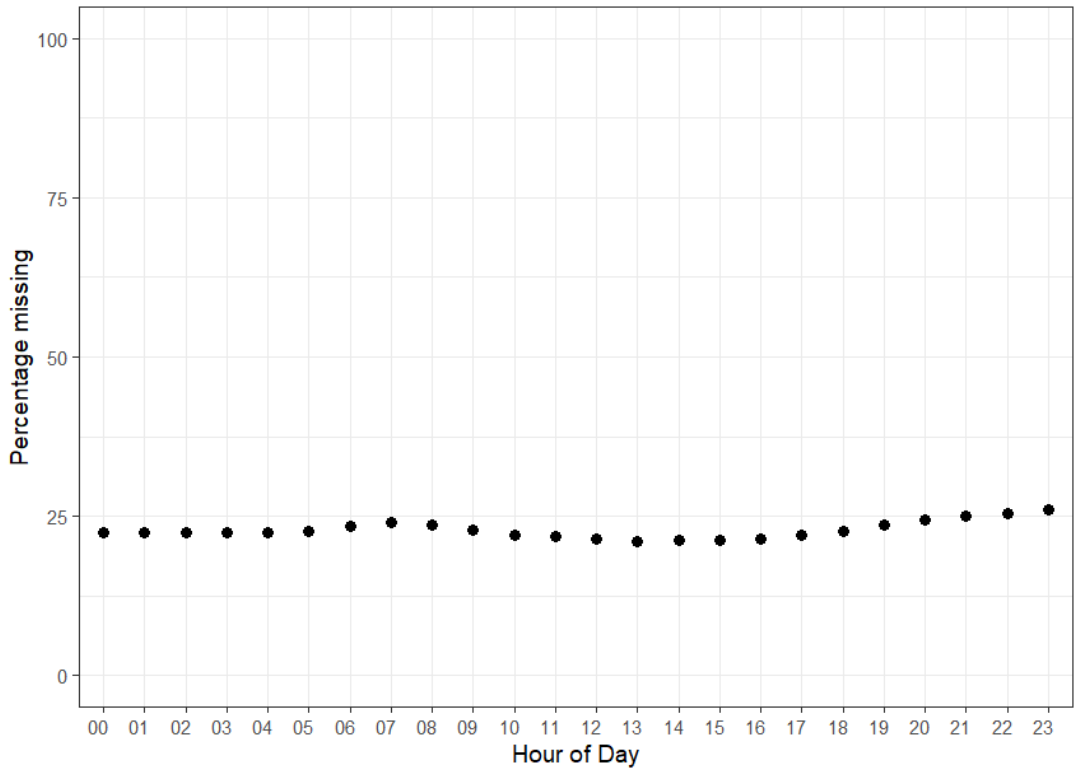



**Figure 6.3** The percentage of missing data for each hour of the day in the NoHoW trial.

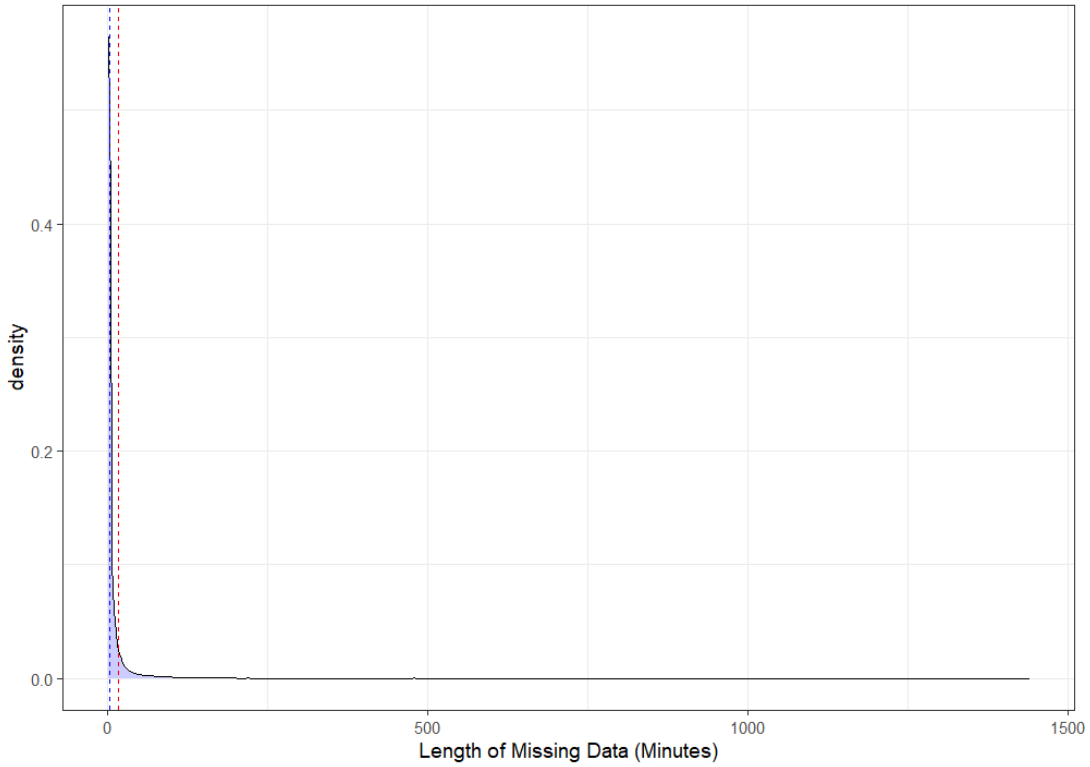



**Figure 6.4** A density plot detailing the lengths of missing data in the NoHoW trial.

Data are presented for missing periods less than 1440 minutes in length. The mean is represented by the red dashed line and the median is represented by the blue dashed line.

### 6.2.7 Imputation methods

Utilising the same simulated missing datasets, the first simulation study tested the methodologies below for dealing with missing data.

#### 6.2.7.1 Removal

The effect of no imputation or adjustment strategy was demonstrated by simply reporting the physical activity summaries for the simulated missing datasets.

#### 6.2.7.2 Mean imputation

Missing data were imputed with the i) mean of all the remaining data and ii) with the mean of the individuals remaining data. This was conducted with the *Hmisc* package in R.

#### 6.2.7.3 Random forest imputation

Random forest imputation was utilised, utilising the 'missForest' package in R. This is a non-parametric imputation method, which implements the original random forest algorithm (Breiman, 2001). Random forest imputation was performed to predict the missing values for steps, heart rate and

calories on each participants data using weekday and hour as observed, non-missing variables. Hyperparameters were selected with consideration of computational feasibility; 100 trees were used in each forest, the number of randomly sampled variables at each split was set to the square root of the number of variables and the maximum number of iterations was set to 5.

#### 6.2.7.4 Multiple imputation

We tested multiple imputation with the use of bootstrapping and predictive mean matching utilising i) the entire sample and ii) individual-level data. In the case of the overall model, age, gender and day of the week were covariates, as they have previously been shown to be associated with differential patterns of physical activity (Berkemeyer et al., 2016; Doherty et al., 2017). In the individual models, the hour of the day was used as an additional covariate. An advantage of multiple imputation is the repetition of the imputation process thus attempting to address the uncertainty associated with a single imputation. A total of 5 imputations were used in the overall model, and in the individual level model, 7 imputations were used. Multiple imputation was implemented with the *Hmisc* package in R.

#### 6.2.7.5 Kalman imputation

Lastly, Kalman smoothing imputation using a structural time series model was tested. Kalman imputation was implemented with the *imputeTS* package in R to impute caloric expenditure, steps and heart rate.

### 6.2.8 Simulation study 2

In simulation study 2, the aim was to investigate how the bias introduced by the NoHoW algorithm, Kalman imputation and individual level multiple imputation may vary depending on the quantity and position of missing data. Individual centred approaches were selected as they were the only individualised approaches that were statistically equivalent to the true data across all activity types in simulation study 1. As in the first simulation study, 14-days (20160 minutes) of data were utilised for each participant. Missingness was simulated randomly throughout the day and in all iterations, the maximum length of each insertion was set to 120 minutes. The simulations were split into 10 windows of missingness, where the number of missing periods inserted for each participant increased incrementally with each simulation window. In the first window, the number of missing periods per participant was sampled from a uniform distribution between 0-10, the second between 10-20 up to the tenth which inserted 90-100 missing periods in each iteration. Within each window, 20 simulations

were conducted per participant, for a total of 21,800 iterations of each algorithm overall.

### 6.2.9 Physical activity metrics

Each of the imputation methods tested in both simulation studies were used to address several distinct physical activity metrics including total steps, TDEE and minutes of sedentary, light, moderate and vigorous physical activity. Both steps and TDEE for a given interval are extracted from the FB and time in each of sedentary, light, moderate and vigorous are defined by the heart rate reserve (HRR) method which is computed for each minute in the dataset. To facilitate this method, maximum heart rate was estimated for each participant using the Tanaka method; (208 - 0.7 x age) (Tanaka et al., 2001). To define resting heart rate, the sleeping heart rate was determined, which was defined as the mean of the lowest 20 consecutive minutes observed between 00:00 and 08:00 am, when steps/min were < 5. After sleeping heart rate was defined, an 8% increase was used because this represents an estimate of the difference between resting and sleeping heart rate (Kräuchi & Wirz-Justice, 2001). The relative intensity of each minute was then calculated:

$$\%\, HRR = \frac{(HR\ - HR_{REST}\ )}{(\ HR_{MAX} - HR_{REST})} \times\ 100$$

The following cut points were applied: Sedentary (<20% HRR), light (20–40% HRR), moderate (40–60% HRR), and vigorous (≥60% HRR) (Schrack et al., 2018). For each missing minute in the dataset, each of the imputation methods described above were used to impute or scale steps, caloric expenditure and heart rate to produce hourly, daily and average physical activity estimates.

### 6.2.10 Statistical analysis

All data are presented as means and standard deviations unless otherwise stated and a flowchart detailing both simulation studies is available in figure 6.5. To evaluate the performance of each method, RMSE was calculated for all physical activity metrics for hourly, daily and 14-day averages, relative to the observed data. Equivalence tests were performed to investigate whether the models were statistically equivalent to the true data. These methods are further explained in **section 3.4.5**. Statistical analyses were conducted with R version 3.6.3 using a p-value of < 0.05 to determine statistical significance.

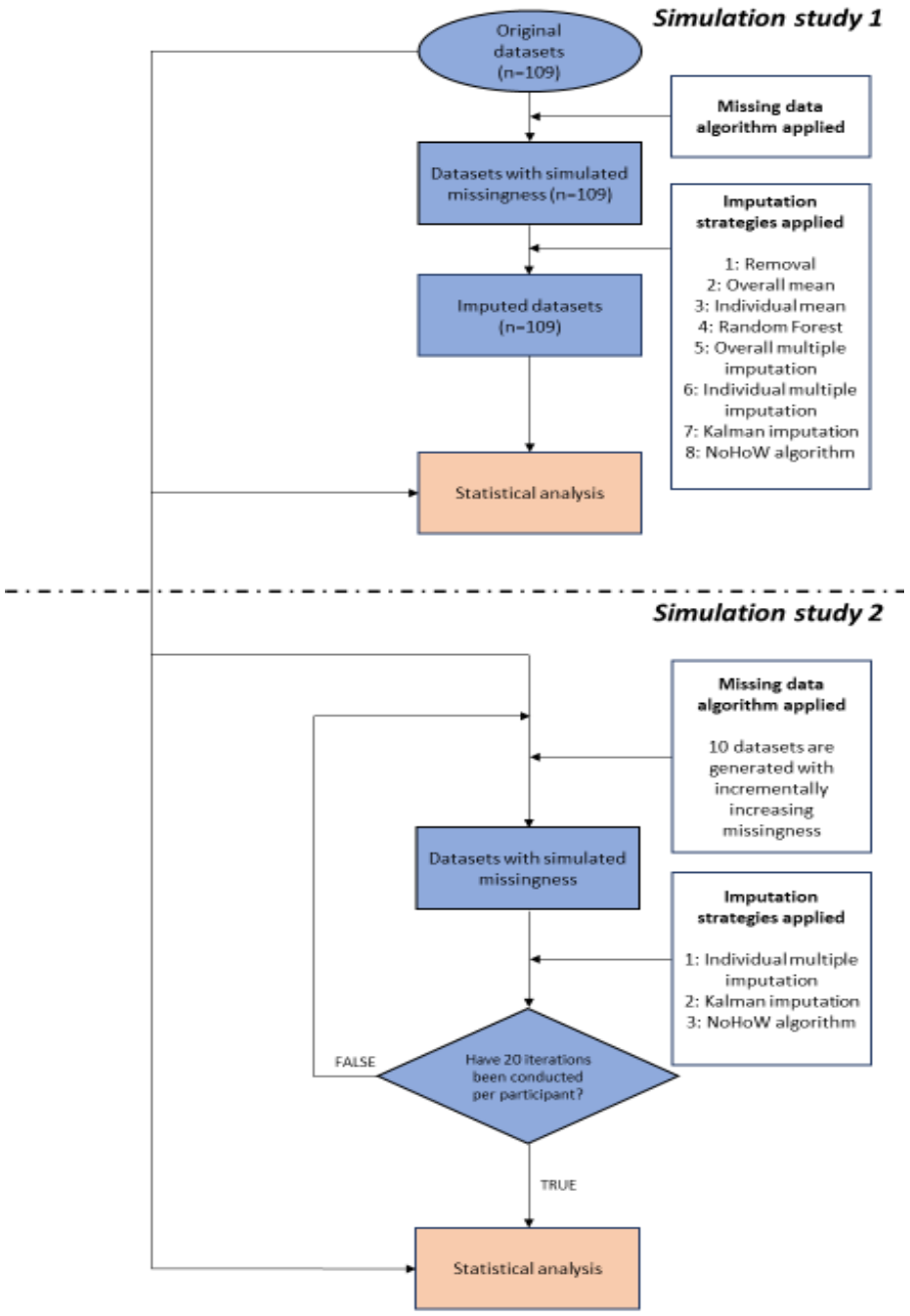



**Figure 6.5** A flowchart detailing the simulation procedures conducted in this study.

## 6.3 Results

The participants meeting the minimum criteria were predominantly female (n= 93, male = 16) and were primarily from the Danish centre (DK = 69, UK

= 23, Portugal = 17), table 6.1 presents the demographic and physical activity results for the included sample. The computation time for each of the included algorithms in the first simulation was as follows: Overall mean imputation: 18.23 Minutes, Individual mean imputation: 1.27 Minutes, Overall multiple imputation: 17.61 Hours, Individual multiple imputation: 17.04 Minutes, Random forest imputation: 4.36 Hours, Kalman imputation: 2.16 Minutes, NoHoW method: 2.12 Seconds.

Table 6.2 illustrates the results of the first simulation study for 14-day, daily and hourly comparisons and table 6.3 presents the results of equivalence tests for each of the methods. For TDEE, Individual multiple imputation had the smallest RMSE for 14-day (36.32 kcal), followed by the NoHoW method (39.51 kcal), and for the hourly comparison, Kalman imputation was superior (14.11 kcal). In the daily comparison, the smallest RMSE was observed for the NoHoW method (115.86 kcal). All methods except removal (mean difference: -343.44 kcal) were statistically equivalent to the true data, with the smallest mean difference observed for Individual multiple imputation. For steps, the lowest RMSE was observed for the NoHoW method for 14-day (397.83 steps) and daily comparison (1366.92 steps) and Kalman imputation for hourly comparison (173.78 steps). All methods except removal (mean difference: -1320.74 steps, p-value >0.05), were statistically equivalent to the true data. In the HRR analysis, multiple imputation methods, Kalman imputation and the NoHoW algorithm were statistically equivalent for all sedentary, light, moderate and vigorous comparisons.

**Table 6.1** Demographic data and physical activity averages for the included sample (n=109).

Total daily EE (TDEE) is presented is kcals/day, sedentary, light, moderate and vigorous are presented in minutes/day.

| | **Mean ± SD** | **Minimum** | **Maximum** |
|---|---|---|---|
| **Age (years)** | 47.46 ± 9.62 | 22 | 75 |
| **Height (M)** | 1.69 ± 0.08 | 1.54 | 1.87 |
| **Weight (kg)** | 84.76 ± 15.59 | 50.5 | 148.4 |
| **BMI (kg/m$^2$)** | 29.64 ± 5 | 20.2 | 44.8 |
| | | | |
| **TDEE (Kcal/day)** | 2626.59 ± 504.66 | 1754.24 | 4492.25 |
| **Steps (Steps/day)** | 10570.34 ± 3208.67 | 3202.50 | 19941.07 |
| **Sedentary (Mins/day)** | 1087.76 ± 112.72 | 847.21 | 1284.64 |
| **Light (Mins/day)** | 266.77 ± 94.83 | 102.29 | 484.14 |
| **Moderate (Mins/day)** | 50.24 ± 31.6 | 6.43 | 132.86 |
| **Vigorous (Mins/day)** | 7.29 ± 9.09 | 0.00 | 47.07 |

**Table 6.2** Mean ± standard deviation estimates for each of the imputation methods tested in simulation study 1.

Total daily EE (TDEE) is presented is kcals, sedentary, light, moderate and vigorous are presented in minutes.

| | | True | Removal | Overall mean | Individual mean | Overall Multiple | Individual Multiple | Random Forest | Kalman | NoHoW |
|---|---|---|---|---|---|---|---|---|---|---|
| **TDEE** | **14-day** | 2626.59 ± 504.66 | 2283.15 ± 445.78 | 2645.66 ± 443.93 | 2645.49 ± 515.87 | 2649.59 ±457.11 | 2638.06 ± 513.24 | 2658.48 ± 571.72 | 2660.96 ± 518.63 | 2653.61 ± 515.8 |
| | **Day** | 2626.59 ±607.05 | 2283.15 ± 583.3 | 2645.66 ± 545.14 | 2645.49 ± 602.91 | 2649.59 ± 555.65 | 2638.06 ± 600.29 | 2658.48 ± 654.09 | 2660.96 ±632.27 | 2653.61 ± 627.21 |
| | **hour** | 109.7 ± 65.24 | 100.12 ± 65.22 | 110.23 ± 61.53 | 110.25 ± 62.03 | 110.36 ± 61.7 | 110.01 ± 62.31 | 110.61 ± 63.2 | 110.69 ± 66.86 | 110.61 ± 67.58 |
| **Steps** | **14-day** | 10570.34 ± 3208.67 | 9249.6 ± 2867.46 | 10718.22 ± 2860.93 | 10716.67 ± 3309.78 | 10741.09 ± 2817.02 | 10593.71 ± 3274.98 | 10049.5 ± 3472.95 | 10755.97 ± 3249.79 | 10791.34 ± 3309.09 |
| | **Day** | 10570.34 + 4775.05 | 9249.6 ± 4447 | 10718.22 ± 4360.78 | 10716.67 ± 4657.6 | 10741.09 ± 4343.52 | 10593.71 ± 4637.29 | 10049.5 ± 4804.18 | 10755.97 ± 5013.85 | 10791.34 ± 5009.86 |
| | **hour** | 442.55 ± 738.52 | 405.61 ± 699.45 | 446.55 ± 694.48 | 446.58 ± 696.27 | 447.17 ± 695.47 | 442.93 ± 697.7 | 428.19 ± 700.15 | 448.23 ± 751.38 | 449.94 ± 763.31 |
| **Sedentary** | **14-day** | 1087.76 ± 112.72 | 956.05 ± 101.01 | 1139.11 ± 109.98 | 1151.53 ± 105.13 | 1118.75 ± 100.63 | 1120.96 ± 110.35 | 1138.12 ± 112.52 | 1101.93 ± 117.73 | 1105.57 ± 116.03 |
| | **Day** | 1087.76 ± 170.63 | 956.05 ± 174.15 | 1139.11 ± 160.95 | 1151.53 ± 156.73 | 1118.75 ± 152.3 | 1120.96 ± 157.97 | 1138.12 ± 165.97 | 1101.93 ± 175.74 | 1105.57 ± 173.68 |

| | | True | Removal | Overall mean | Individual mean | Overall Multiple | Individual Multiple | Random Forest | Kalman | NoHoW |
|---|---|---|---|---|---|---|---|---|---|---|
| | **hour** | 45.51 ± 17.82 | 41.93 ± 19.1 | 47.03 ± 16.83 | 47.38 ± 16.61 | 46.46 ± 16.44 | 46.54 ± 16.52 | 47.01 ± 17 | 45.99 ± 18.04 | 46.07 ± 17.89 |
| **Light** | **14-day** | 266.77 ± 94.83 | 235.5 ± 84.53 | 247.25 ± 93.15 | 236.96 ± 86.51 | 264.11 ± 84.32 | 261.72 ± 91.8 | 247.05 ± 94.01 | 269.32 ± 96.42 | 274.66 ± 96.81 |
| | **Day** | 266.77 ± 139.25 | 235.5 ± 127.82 | 247.25 ± 134.98 | 236.96 ± 129.6 | 264.11 ± 126.65 | 261.72 ± 131.03 | 247.05 ± 137.41 | 269.32 ± 143.71 | 274.66 ± 144.45 |
| | **hour** | 11.17 ± 14.15 | 10.33 ± 13.59 | 10.65 ± 13.86 | 10.36 ± 13.62 | 11.11 ± 13.47 | 11.03 ± 13.54 | 10.63 ± 13.92 | 11.2 ± 14.54 | 11.44 ± 14.69 |
| **Moderate** | **14-day** | 50.24 ± 31.6 | 44.63 ± 28.42 | 44.63 ± 28.42 | 44.63 ± 28.42 | 48.81 ± 28.63 | 48.25 ± 31.18 | 44.65 ± 28.43 | 48.76 ± 31.17 | 52.22 ± 33.37 |
| | **Day** | 50.24 ± 47.85 | 44.63 ± 43.95 | 44.63 ± 43.95 | 44.63 ± 43.95 | 48.81 ± 44.07 | 48.25 ± 45.54 | 44.65 ± 43.95 | 48.76 ± 49.15 | 52.22 ± 51.02 |
| | **hour** | 2.11 ± 5.53 | 1.96 ± 5.29 | 1.96 ± 5.29 | 1.96 ± 5.29 | 2.07 ± 5.29 | 2.06 ± 5.3 | 1.96 ± 5.29 | 2.07 ± 5.67 | 2.18 ± 5.88 |
| **Vigorous** | **14-day** | 7.29 ± 9.09 | 6.51 ± 8.38 | 6.51 ± 8.38 | 6.51 ± 8.38 | 6.92 ± 8.35 | 6.86 ± 9.13 | 6.51 ± 8.38 | 7.17 ± 9.03 | 7.55 ± 9.57 |
| | **Day** | 7.29 ± 15.48 | 6.51 ± 14.61 | 6.51 ± 14.61 | 6.51 ± 14.61 | 6.92 ± 14.59 | 6.86 ± 14.99 | 6.51 ± 14.61 | 7.17 ± 15.66 | 7.55 ± 16.62 |
| | **hour** | 0.3 ± 2.45 | 0.29 ± 2.36 | 0.29 ± 2.36 | 0.29 ± 2.36 | 0.3 ± 2.36 | 0.3 ± 2.37 | 0.29 ± 2.36 | 0.31 ± 2.52 | 0.32 ± 2.6 |

**Table 6.3** Mean ± standard deviation estimates and equivalence test results for each of the imputation methods tested in simulation study 1.

Total daily EE (TDEE) is presented is kcals. Sedentary, light, moderate and vigorous are presented in minutes.

| | | True | Imputed | Mean difference | Bounds | P-value lower | P-value upper |
|---|---|---|---|---|---|---|---|
| **TDEE** | **Removal** | 2626.59 ± 504.66 | 2283.15 ± 445.78 | -343.44 | ± 262.66 | 1 | 0 |
| | **Overall mean** | 2626.59 ± 504.66 | 2645.66 ± 443.93 | 19.08 | ± 262.66 | 0 | 0 |
| | **Individual mean** | 2626.59 ± 504.66 | 2645.49 ± 515.87 | 18.9 | ± 262.66 | 0 | 0 |
| | **Overall Multiple** | 2626.59 ± 504.66 | 2649.59 ± 457.11 | 23.01 | ± 262.66 | 0 | 0 |
| | **Individual Multiple** | 2626.59 ± 504.66 | 2638.06 ± 513.24 | 11.48 | ± 262.66 | 0 | 0 |
| | **Random Forest** | 2626.59 ± 504.66 | 2658.48 ± 571.72 | 31.89 | ± 262.66 | 0 | 0 |
| | **Kalman** | 2626.59 ± 504.66 | 2660.96 ± 518.63 | 34.37 | ± 262.66 | 0 | 0 |
| | **NoHoW** | 2626.59 ± 504.66 | 2653.61 ± 515.8 | 27.02 | ± 262.66 | 0 | 0 |
| **Steps** | **Removal** | 10570.34 ± 3208.67 | 9249.6 ± 2867.46 | -1320.74 | ± 1057.03 | 1 | 0 |
| | **Overall mean** | 10570.34 ± 3208.67 | 10718.22 ± 2860.93 | 147.88 | ± 1057.03 | 0 | 0 |
| | **Individual mean** | 10570.34 ± 3208.67 | 10716.67 ± 3309.78 | 146.33 | ± 1057.03 | 0 | 0 |
| | **Overall Multiple** | 10570.34 ± 3208.67 | 10741.09 ± 2817.02 | 170.75 | ± 1057.03 | 0 | 0 |
| | **Individual Multiple** | 10570.34 ± 3208.67 | 10593.71 ± 3274.98 | 23.37 | ± 1057.03 | 0 | 0 |
| | **Random Forest** | 10570.34 ± 3208.67 | 10049.5 ± 3472.95 | -520.84 | ± 1057.03 | 0 | 0 |
| | **Kalman** | 10570.34 ± 3208.67 | 10755.97 ± 3249.79 | 185.63 | ± 1057.03 | 0 | 0 |
| | **NoHoW** | 10570.34 ± 3208.67 | 10791.34 ± 3309.09 | 221 | ± 1057.03 | 0 | 0 |
| **Sedentary** | **Removal** | 1087.76 ± 112.72 | 956.05 ± 101.01 | -131.71 | ± 108.78 | 1 | 0 |

| | | True | Imputed | Mean difference | Bounds | P-value lower | P-value upper |
|---|---|---|---|---|---|---|---|
| | **Overall mean** | 1087.76 ± 112.72 | 1139.11 ± 109.98 | 51.36 | ± 108.78 | 0 | 0 |
| | **Individual mean** | 1087.76 ± 112.72 | 1151.53 ± 105.13 | 63.78 | ± 108.78 | 0 | 0 |
| | **Overall Multiple** | 1087.76 ± 112.72 | 1118.75 ± 100.63 | 30.99 | ± 108.78 | 0 | 0 |
| | **Individual Multiple** | 1087.76 ± 112.72 | 1120.96 ± 110.35 | 33.2 | ± 108.78 | 0 | 0 |
| | **Random Forest** | 1087.76 ± 112.72 | 1138.12 ± 112.52 | 50.36 | ± 108.78 | 0 | 0 |
| | **Kalman** | 1087.76 ± 112.72 | 1101.93 ± 117.73 | 14.17 | ± 108.78 | 0 | 0 |
| | **NoHoW** | 1087.76 ± 112.72 | 1105.57 ± 116.03 | 17.81 | ± 108.78 | 0 | 0 |
| **Light** | **Removal** | 266.77 ± 94.83 | 235.5 ± 84.53 | -31.27 | ± 26.68 | 1 | 0 |
| | **Overall mean** | 266.77 ± 94.83 | 247.25 ± 93.15 | -19.51 | ± 26.68 | 0.046 | 0 |
| | **Individual mean** | 266.77 ± 94.83 | 236.96 ± 86.51 | -29.8 | ± 26.68 | 0.949 | 0 |
| | **Overall Multiple** | 266.77 ± 94.83 | 264.11 ± 84.32 | -2.65 | ± 26.68 | 0 | 0 |
| | **Individual Multiple** | 266.77 ± 94.83 | 261.72 ± 91.8 | -5.05 | ± 26.68 | 0 | 0 |
| | **Random Forest** | 266.77 ± 94.83 | 247.05 ± 94.01 | -19.72 | ± 26.68 | 0.001 | 0 |
| | **Kalman** | 266.77 ± 94.83 | 269.32 ± 96.42 | 2.55 | ± 26.68 | 0 | 0 |
| | **NoHoW** | 266.77 ± 94.83 | 274.66 ± 96.81 | 7.89 | ± 26.68 | 0 | 0 |
| **Moderate** | **Removal** | 50.24 ± 31.6 | 44.63 ± 28.42 | -5.61 | ± 5.02 | 0.938 | 0 |
| | **Overall mean** | 50.24 ± 31.6 | 44.63 ± 28.42 | -5.61 | ± 5.02 | 0.938 | 0 |
| | **Individual mean** | 50.24 ± 31.6 | 44.63 ± 28.42 | -5.61 | ± 5.02 | 0.938 | 0 |
| | **Overall Multiple** | 50.24 ± 31.6 | 48.81 ± 28.63 | -1.44 | ± 5.02 | 0 | 0 |
| | **Individual Multiple** | 50.24 ± 31.6 | 48.25 ± 31.18 | -1.99 | ± 5.02 | 0 | 0 |

| | | True | Imputed | Mean difference | Bounds | P-value lower | P-value upper |
|---|---|---|---|---|---|---|---|
| | **Random Forest** | 50.24 ± 31.6 | 44.65 ± 28.43 | -5.59 | ± 5.02 | 0.933 | 0 |
| | **Kalman** | 50.24 ± 31.6 | 48.76 ± 31.17 | -1.48 | ± 5.02 | 0 | 0 |
| | **NoHoW** | 50.24 ± 31.6 | 52.22 ± 33.37 | 1.98 | ± 5.02 | 0 | 0 |
| **Vigorous** | **Removal** | 7.29 ± 9.09 | 6.51 ± 8.38 | -0.78 | ± 0.73 | 0.672 | 0 |
| | **Overall mean** | 7.29 ± 9.09 | 6.51 ± 8.38 | -0.78 | ± 0.73 | 0.672 | 0 |
| | **Individual mean** | 7.29 ± 9.09 | 6.51 ± 8.38 | -0.78 | ± 0.73 | 0.672 | 0 |
| | **Overall Multiple** | 7.29 ± 9.09 | 6.92 ± 8.35 | -0.37 | ± 0.73 | 0.002 | 0 |
| | **Individual Multiple** | 7.29 ± 9.09 | 6.86 ± 9.13 | -0.43 | ± 0.73 | 0.005 | 0 |
| | **Random Forest** | 7.29 ± 9.09 | 6.51 ± 8.38 | -0.78 | ± 0.73 | 0.672 | 0 |
| | **Kalman** | 7.29 ± 9.09 | 7.17 ± 9.03 | -0.12 | ± 0.73 | 0 | 0 |
| | **NoHoW** | 7.29 ± 9.09 | 7.55 ± 9.57 | 0.26 | ± 0.73 | 0 | 0 |

In the second simulation study, which is visually represented as boxplots in Figure 6.6, the aggregated RMSE for each of the tested approaches tended to increase with the proportion of missing data. For the TDEE estimation (Figure 6.6a), the first iteration (1% missingness added) resulted in a mean RMSE of 31.14 kcal/day for the NoHoW method (range 28.82 – 33.12 kcal/day) compared to multiple imputation: 21.30 kcal/day (range 19.20 – 23.11 kcal/day) and Kalman imputation: 37.44 kcal/day (range 35.49– 39.90 kcal/day). Comparatively, at the 10th insertion of missingness (~28% missingness added) a maximum RMSE of 68.89 kcal/day, 68.05 kcal/day and 72.55 kcal/day was observed for NoHoW, multiple imputation and Kalman imputation, respectively. For steps (Figure 6.6b), evidence of slightly superior performance was observed for multiple imputation at the lower levels of missingness (<19%). However, mean RMSE values for each of the methods remained similar and did not differ by more than 86 steps/day. In the HRR analysis, differences were the greatest in the sedentary comparison (Figure 6.6c), with the NoHoW and Kalman methods having a lower mean RMSE than multiple imputation at each window. The largest difference was observed at 28% missingness, where the mean RMSE values were 24.87 mins/day (range: 23.15 – 26.39 mins/day) for the NoHoW method, 55.56 (range 53.69- 57.76) mins/day for multiple imputation and 23.73 mins/day (range 21.46- 26.89 mins/day) for Kalman imputation. For light (Figure 6.6d) and moderate (Figure 6.6e) the NoHoW method showed the lowest mean RMSE values after 13% missingness. Its largest mean RMSE of 15.19 mins/day (range 12.81- 17.42 mins/day) for light activity and 5.38 mins/day (range 4.72- 6.26 mins/day) for moderate activity were observed at 28 % missingness. Lastly, in the vigorous activity simulation (Figure 6.6f), multiple imputation had the lowest mean RMSE with <7% added missingness but Kalman and NoHoW methods were superior at higher levels of missingness. In the 28 % missingness window, NoHoW reached a mean RMSE of 2.25 mins/day (range 1.84 – 3.03 mins/day) mins/day and Kalman reached 2.28 mins/day (range 1.85-2.95 mins/day). An extensive table of results from the second simulation study is available in appendix 2.1.

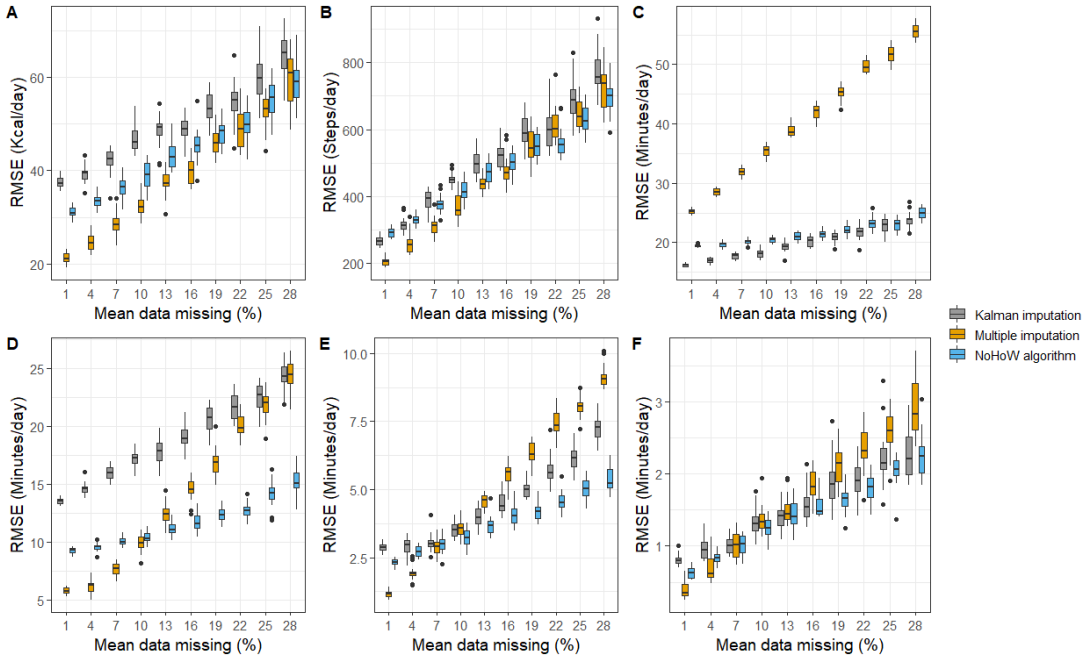



**Figure 6.6** Boxplots detailing root mean squared error (RMSE) values from simulation study 2 for each window of missingness.

Data are presented for TDEE (A), Steps (B), Sedentary (C), Light (D), Moderate (E), Vigorous (F). Mean missing data refers to the additional data added in the simulations.

## 6.4 Discussion

The use of commercial activity monitors in research environments is proliferating, creating new research opportunities within the fields of energy balance or physical activity. It is necessary to take steps to ensure the integrity of these data is not challenged by missing data. The purpose of the present study was to develop and test a methodology to account for missingness in EE/physical activity data collected with a commercial activity monitor in a free-living environment. In the first set of experiments, ICC analyses were used to show that if data are scaled within an hour, the relative data requirements to meet an ICC threshold of 0.9 are minimal (~5 minutes). This relates to the relative similarity between 'local' data points, as confirmed by the autocorrelation analyses. It is shown that if the data are not scaled by wear time the relative requirements for a day equates to approximately 18 hours per day. This is in contrast to a previous study, which showed that relative to a 14 hours/day criterion, at least 13 hours/day of accelerometer data are required (Herrmann et al., 2013). This slight discrepancy in the proportion of the day required may relate to the inclusion

of night hours in this sample. Given the likelihood that this is a highly sedentary period, missing data at night is likely to be less influential on daily totals.

In simulation study 1, each of the tested methods were used to impute metrics that are likely to be of importance depending on the specific research aims. The results suggest different outcomes depending on the metric selected, for instance, random forest imputation, overall mean and individual mean methods did not impute vigorous or moderate minutes regularly, as reflected in the non-significant equivalent results (indicating these methods are not statistically equivalent). This is likely due to the low proportion of the day in which these activities are performed. In the first simulation study, a slight tendency for the NoHoW method to overestimate minutes of moderate and vigorous activity was observed. This may relate to the position of the missing data in simulation 1; For example, if missing data occurs in the sedentary period after an exercise bout then this period will be overestimated. As exercise is infrequent in non-athlete populations this is unlikely to result in a large error in mean differences. Indeed, the estimates for moderate and vigorous differed by < 2 minutes/day in the 14-day comparison. Researchers should consider imputation strategies based on observed activity data from their sample or should select methodologies which are statistically equivalent in the specific activities of interest.

All tested methods resulted in a RMSE which was lower than no imputation (i.e. removal). Making no attempt to adjust for missingness effectively assumes that activity was 0 and the results presented here demonstrate the potential implications of this. In the first experiment, ~14% of the day was missing on average with ~12% inserted, equating to a wear time of 20-22 hours, which falls within the acceptable levels of missingness for most accelerometer research (Choi et al., 2011; Ridgers & Fairclough, 2011) and therefore evidences the importance of using one of these methods even in the case of relatively small quantities of missing data. Of the imputation methods tested, an advantage of individual-centred methods was observed, specifically Kalman imputation, individual multiple imputation and the NoHoW algorithm. Indeed, in the second simulation study, in which the maximal missingness approached double the quantity of the first simulation study the RMSE for TDEE was lower than the values observed for removal, overall mean and random forest imputation in simulation study 1, indicating the efficacy of these methods.

Missingness was simulated approximately evenly throughout the entire 24-hour period because of the observed patterns of missingness in the NoHoW trial. This is contrary to a previous study observing that missing data patterns more frequently occur at the beginning and end of the day (Xu et al., 2018). It is of note that the present study used wrist-worn devices compared to the aforementioned study, which utilised hip worn accelerometers. Unlike wrist-worn monitors, hip-worn accelerometers are generally removed with changing of clothes and sleeping . This may encourage compliance in wrist-worn monitors (Diaz, Krupka, Chang, Shaffer, et al., 2016) and contribute to a more uniform distribution of missingness throughout the day.

The relative computational simplicity of the NoHoW method is a significant advantage. Accelerometer data of this kind can be extremely high volume and researchers must select their imputation strategy with consideration of both error reduction and computational feasibility. It may be possible to utilise advanced machine learning techniques to impute missing data, but these methods are computationally expensive and may be technically inaccessible to many researchers. Also, more information (e.g. physiological, psychological or behavioural factors) may allow for more accurate multivariate imputation techniques but in free-living settings this information is likely to be limited, thus the method presented here is likely to be widely applicable. A further advantage of the present study is the testing of numerous activity metrics in addition to steps. Steps are a highly interpretable and relatable metric produced by wearable devices and some evidence suggests that estimates of steps from Fitbit devices are more valid and reliable than other derived variables, i.e. TDEE (Feehan et al. 2018; O'Driscoll et al. 2020; O'Driscoll et al. 2020) although machine learning techniques may facilitate the refinement of EE estimates (O'Driscoll et al. 2020). Nevertheless, the metric of interest to researchers will vary depending on the aims and hypotheses of a study and the NoHoW method, Kalman imputation and individual level multiple imputation perform particularly well across a variety of physical activity metrics.

### 6.4.1 Limitations

Key limitations of the present study are the utilisation of participants with a high proportion of wear time (>97.5%). Whilst highly adherent participants were required to have a near-complete dataset to validate against, the included participants may be in some way behaviourally different from the participants that remove the FB more frequently. Second, missing data were inserted at random positions, and it remains uncertain how representative

this is of free-living data in other studies. Participants may remove devices for comfort, aesthetic reasons, charging or under conditions where they would not wish to have measurements made (e.g. extreme sedentariness) and thus, it is possible that missingness is not completely at random (Sterne et al., 2009) and may differ between populations and research studies. Unfortunately, no definitive method exists to test if data are missing at random (Lee & Gill, 2018) and many imputation strategies have limited capabilities to overcome this. However, our second simulation study simulates a wide variety of missing patterns in an attempt to identify such biases and worst-case scenarios in the selected methods. Lastly, the sample is made up predominantly of females with overweight/obesity.

## 6.5 Conclusions

Incorporation of activity monitoring devices is a necessary step in improving physical activity and energy balance tracking in research and clinical settings. A simple and accessible methodology has been proposed in this chapter which effectively reduces the bias introduced to physical activity estimates by non-wear time and may improve the validity of research conclusions. Other imputation strategies (i.e. multiple imputation and Kalman imputation) performed comparatively well and importantly, all the methods tested in this study are superior to data removal. Researchers and clinicians utilising commercial activity monitors longitudinally should account for missingness and the algorithm presented in this study offers an approach to this.

Performance for TDEE, steps and heart rate show similar trends for the NoHoW algorithm as these metrics are closely related. Despite this, it must be stated that TDEE is the metric of primary interest in this thesis. In terms of performance, a slight advantage of the NoHoW algorithm was observed as missingness increased. Though the error remained small for all experiments (RMSE < 80kcal/day). Substantial differences exist in the speed of computation, however, with the NoHoW algorithm computing orders of magnitude faster than Kalman or Multiple imputation.

# Chapter 7 – Development and validation of machine learning models to estimate energy expenditure from wearable sensors

## 7.1 Introduction

The case has been made throughout this thesis that activity tracking devices have some significant advantages which make them a potentially interesting tool for use in research environments. For energy balance research specifically, these benefits are certainly offset if the accuracy of EE measures is poor. As indicated in **chapter 4** and **5**, the accuracy of activity trackers varies greatly between devices and activities (O'Driscoll, Turicchi, Beaulieu, et al., 2020; O'Driscoll, Turicchi, Hopkins, et al., 2020; Shcherbina et al., 2017), which limits their use when quantifying energy balance and activity behaviours.

The potential of machine learning techniques to model the complex interactions between accelerometer data, physiological variables and rate of EE has been recognised for some time. An early study showed that an artificial neural network can be trained using accelerometer data as input to predict EE in a whole-body indirect calorimetry chamber (Rothney et al., 2007). Furthermore, Pober et al. utilised quadratic discriminant analysis and hidden Markov models to classify activity and subsequently estimate the proportion of time at different rates of EE (Pober et al., 2006). Research groups have built on these early findings and reported highly accurate algorithms in a variety of activities (Ahmadi et al., 2020; Ellis et al., 2014, 2016; Montoye, Begum, et al., 2017; Staudenmayer et al., 2009). Researchers may take two broad approaches when modelling physical activity; First, attempting to predict the rate of EE as a continuous variable (i.e. METs) or second, classifying a minute as sedentary, light or moderate to vigorous physical activity (MVPA) and both of these approaches are important for health research. Regression approaches could be used to derive total EE for a subject and this estimate can subsequently be incorporated into energy balance models to calculate EI (Shook et al., 2018). Alternatively, accurately determining the time an individual spends in broader categories of activity and/or the intensity of that activity can be important for public health guidance. For example, successful weight maintenance in the National Weight Control Registry and weight

management recommendations are often defined based on the time an individual spends in MVPA (Ostendorf et al., 2018). Machine learning algorithms have the potential to enhance physical activity assessment beyond traditional count-based methods, which despite being more accessible, may not be sufficiently accurate for the assessment of EE and intensity classifications (Lyden et al., 2011).

Recently, it has been demonstrated in a laboratory validation study that accelerometer and physiological sensor outputs can be modelled using random forests to predict the rate of 'steady-state' EE in commercial and research-grade activity monitors, with the accuracy surpassing the proprietary algorithms of the SWA (O'Driscoll, Turicchi, Hopkins, et al. 2020). A limitation of this work was that the number of activities in which EE was measured was limited and the generalisability of these algorithms remains uncertain.

A method for continued refinement of predictive algorithms is to aggregate more than one dataset to provide larger, more diverse training data with more activities (Chowdhury et al., 2017). More data presents a new optimisation problem, which (because of different assumptions made by different algorithms) means there is no guarantee that any algorithm will minimise error on all problems (Wolpert & Macready, 1997). For machine learning models to be used in general health research settings it is critical to evaluate the generalisability of prediction algorithms. The extent to which an algorithm will generalise is influenced by the characteristics of the sample, the activity types as well as the size and quality of the training data. One approach which addresses each of these limitations is to evaluate prediction algorithms on different samples, using data collected under different conditions. In addition to generalisability, a combination of variable datasets collected under different experimental conditions (Farrahi et al., 2020) may help to increase the accuracy of predictions.

### 7.1.1 Chapter aims

This study aggregates two distinct datasets of concurrent inputs from multiple wearable devices and measured EE (indirect calorimetry). The primary aims were to develop and evaluate classification and regression algorithms to i) predict the rate of EE and ii) classify a single minute as sedentary, light or MVPA. Algorithms were validated using leave-one-subject-out cross-validation (LOSO) as well as an out-of-sample validation. Concurrently, the SWA is evaluated as this is a device which has been shown to outperform accelerometer-based monitors when classifying activity

minutes (Calabró et al., 2014) and is one of the most accurate wrist or arm-based monitors for estimating EE (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020).

## 7.2 Methods

### 7.2.1 Studies and protocols

The present study aggregated data collected as part of two separate studies at The Human Appetite Research Unit, University of Leeds. Both studies are described in **chapter 3** (See device validation study in **section 3.1.1, 3.2.1** and **3.3.1** and 'TEED study' in **sections 3.1.2, 3.2.2** and **3.3.2**). Participant information for both samples is shown in table 7.1. On average, the sample in study 2 (TEED study) had proportionately more males, a lower age, a lower average percentage of FM and a higher RMR, compared to study 1 (Device validation study).

**Table 7.1** Characteristics of the included sample.

Data are presented as mean ± SD. Abbreviations: FFM = Fat free mass, FM = Fat mass, RMR = Resting metabolic rate. Weight refers to the weight measured at visit 1 for exercise testing, whereas body composition was collected at a later date. Body composition was not available for all participants in study 1.

| Study | N (Female) | Age (years) | Height (cm) | Weight (kg) | FFM (kg) | FM (kg) | FM (%) | RMR (kcal/d) |
|---|---|---|---|---|---|---|---|---|
| **1** | 59 (41) | 44.4 ± 14.1 | 167.5 ± 8.9 | 75.7 ± 13.6 | 49.8 ± 8.9 | 24.8 ± 10.7 | 32.5 ± 10.3 | 1581.8 ± 280.4 |
| **2** | 30 (13) | 31.9 ± 10.2 | 171.9 ± 9.2 | 70.6 ± 12.9 | 55 ± 12.6 | 15.1 ± 7.1 | 21.7 ± 8.7 | 1769.3 ± 435.8 |

#### 7.2.1.1 Study 1

Full details of study 1 have been published previously (O'Driscoll, Turicchi, Hopkins, et al. 2020). The protocol of study 1 consisted of 10 activities, each performed for 5 minutes in the following set order: sitting, standing, treadmill walking and incline walking (4 km/h), jogging and incline jogging (6–8 km/h). Participants then rested for 3 minutes and transitioned to a cycle ergometer for low and moderate-intensity cycling. After another period of recovery, participants performed a folding task and a sweeping task. Due to variation

in physical fitness, the jogging task (n = 49), incline jogging (n = 30) and the moderate cycling tasks (n = 58) were not performed by all participants.

#### 7.2.1.2 Study 2

In study 2 (TEED study), participants visited the lab having refrained from eating or consuming caffeine for at least 4 hours. Weight and height were obtained from a SECA 704s stadiometer and electronic scale (SECA, Germany) and subsequently an activity protocol was performed. All activities were performed in 5-minute increments and the order was identical for all participants. Firstly, resting tasks were performed where participants lay supine, then sat in a backed chair and then stood. Next, after a 2-minute unstructured transitional period, participants performed seated typing, standing ironing and wiping surfaces whilst standing. After another 2-minute transition, participants walked on a treadmill at 4 km/h, walked at an incline of 5% at 4 km/h and subsequently jogged at 7 km/h. Participants then rested for 10 minutes. After the unstructured resting period, participants performed low-intensity and moderate-intensity cycling, low-intensity and moderate-intensity rowing and low-intensity and moderate-intensity cross-training (elliptical), with 1-minute transitions between each. The intensity was determined by self-selected perceived exertion. In study 2, one participant did not perform rowing or elliptical tasks.

### 7.2.2 Body composition assessment

In both studies, body composition was estimated using ADP (BodPod, Life Measurement, Inc.; USA), n=57 in study 1 and n=30 in study 2, by the method described in **section 3.4.1.5.**

### 7.2.3 Energy expenditure

In both studies, RMR was determined by the methods described in **section 3.4.2.2**. If RMR data were unavailable (n=3 across both studies) RMR was approximated with BMI specific equations (Müller et al., 2004). During the activity sessions, minute-level EE was obtained from a stationary metabolic cart (Vyntus CPX, Jaeger-CareFusion, UK) which is discussed further in **section 3.4.2.3** and these data were expressed relative to each subject's measured RMR to derive METs, which served to eliminate the proportion of EE attributable to RMR.

### 7.2.4 Devices

Accelerometer and physiological data were collected by various sensors throughout both protocols. The Polar H7 chest strap (Polar Electro,

Kempele, Finland) was used to measure heart rate. An ActiGraph GT3-X accelerometer (AG; Actigraph, Pensacola, FL, USA) and a Fitbit charge 2 (FB; Fitbit Inc, San Francisco, CA, USA) were attached securely on the non-dominant wrist. Participants also wore the SenseWear Armband Mini (SWA; BodyMedia Inc., Pittsburgh, PA, USA) on the upper arm. The devices used in this study are all detailed in **section 3.5.2**.

### 7.2.5 Data aggregation

The sensor outputs were obtained from the device-specific software, aggregated to the minute-level and time-matched to the criterion EE data. Data loss attributable to device malfunction were as follows: In study 1, two participant's FB data, one participant's AG data and one participant's polar heart rate data were lost. In study 2, one SWA and one FB dataset were lost due to device failure. Given the slightly different data availability in each model, the results report the number of minutes used and the number of participants. All minutes in which EE data were available (i.e. face mask was not removed) were included in this analysis, and the aggregation of the datasets by time was conducted in Python 3.7.6 and R version 3.6.3.

For activity-specific analyses, activities were grouped into broader categories. Specifically, 'Activities of daily living' (ADL), which involved folding, sweeping, typing, ironing and wiping surfaces. Distinct categories were assigned for 'Cycling', 'Elliptical', 'Rowing', 'Running' and 'Walking'. The 'Sedentary' activities involved all sitting, standing and supine tasks. The 'Transitional' category refers to unstructured resting or transitional minutes.

### 7.2.6 Model features

Predictive models were built for FB, AG and SWA and the features used in each model are detailed in table 7.2. Each device used a combination of subject-level features, accelerometer features and physiological features, which have all been related to the rate of EE in previous literature (Brage et al., 2005; Ceesay et al., 1989; O'Driscoll, Turicchi, Beaulieu, et al., 2020; Rothney et al., 2007; Whybrow, Ritz, Horgan, & Stubbs, 2013). The exact features varied depending on features available in each device. Where small (<5 minutes) heart rate gaps existed (e.g. loss of signal between the respective heart rate sensor and the skin), linear interpolation was used to fill gaps. As activity in the preceding minutes influences the rate of EE at the measurement point (McArdle et al., 2010, pp 192 - 234), several time-lagged features were computed; for steps (FB and SWA), vector magnitude (AG), FB heart rate (FB) and polar heart rate (SWA and AG) the change, for t-1

minutes through to t-5 minutes were included as predictive features. Also, the mean and standard deviation of the last 5 minutes were used as predictive features. If time-lagged variables could not be computed due to data missing (i.e. for the first minutes for each subject), backwards imputation was performed using the next available observation.

As a constant variance is a central requirement of some of the algorithms tested in this study, all numeric features were standardised before training by the formula $z = \frac{(x-\mu)}{sd}$ where $\mu$ and $sd$ refer to the variable mean and standard deviation, respectively.

**Table 7.2** Predictive features used in each of the models.

For each device, the subject characteristics, acceleration features and physiological features are listed.

| Device | Category | Features |
|---|---|---|
| **FB** | **Subject features** | Gender, age, height, weight, sitting heart rate |
| | **Acceleration features** | **Steps:** |
| | | Steps mean, |
| | | Steps difference (t-1, t-2, t-3, t-4, t-5 minutes), |
| | | Steps mean of last 5 minutes, |
| | | Steps standard deviation of last 5 minutes |
| | **Physiological features** | **Fitbit heart rate features:** |
| | | Fitbit heart rate above sitting heart rate, |
| | | Fitbit heart rate percentage of maximum heart rate, Fitbit heart rate mean, |
| | | Fitbit heart rate difference (t-1, t-2, t-3, t-4, t-5), |
| | | Fitbit heart rate mean of last 5 minutes, |
| | | Fitbit heart rate standard deviation of last 5 minutes |
| **AG** | **Subject features** | Gender, age, height, weight |
| | **Acceleration features** | **X,Y,Z Features:** |
| | | Minimum, Maximum, Mean, Standard Deviation, |
| | | Median Crossings, |
| | | $10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, $90^{th}$ Percentiles, |
| | | Correlations (XY, XZ, YZ), |
| | | Dominant frequency, |
| | | Dominant frequency magnitude |
| | | **First order differential of X,Y,Z Features:** |
| | | Minimum, Maximum, Mean, Standard Deviation, |
| | | Median Crossings, |
| | | $10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$, $90^{th}$ Percentiles, |

| Device | Category | Features |
|---|---|---|
| | | Correlations (XY, XZ, YZ), |
| | | Dominant frequency, |
| | | Dominant frequency magnitude |
| | | **Vector magnitude:** |
| | | Vector magnitude mean, |
| | | Vector magnitude difference (t-1, t-2, t-3, t-4, t-5 Minutes), Vector magnitude mean of last 5 minutes, |
| | | Vector magnitude standard deviation of last 5 minutes |
| | **Physiological features** | **Polar heart rate features:** |
| | | Polar heart rate above sitting heart rate, |
| | | Polar heart rate percentage of maximum heart rate, Polar heart rate mean, |
| | | Polar heart rate difference (t-1, t-2, t-3, t-4, t-5 Minutes), |
| | | Polar heart rate mean of last 5 minutes, |
| | | Polar heart rate standard deviation of last 5 minutes |
| **SWA** | **Subject features** | Gender, age, height, weight |
| | **Acceleration features** | **X, Y, Z Features:** |
| | | Peaks, Mean of absolute deviation, Average |
| | | **Steps:** |
| | | Steps mean, Steps difference (t-1, t-2, t-3, t-4, t-5 Minutes), Steps mean of last 5 minutes, |
| | | Steps standard deviation of last 5 minutes |
| | **Physiological features** | **Polar heart rate features:** |
| | | Polar heart rate above sitting heart rate, |
| | | Polar heart rate percentage of maximum heart rate, Polar heart rate mean, |
| | | Polar heart rate difference (t-1, t-2, t-3, t-4, t-5 Minutes), |
| | | Polar heart rate mean of last 5 minutes, |
| | | Polar heart rate standard deviation of last 5 minutes |
| | | **SenseWear sensors:** |
| | | Near body temperature average, Galvanic skin response average, Skin temperature average |

### 7.2.7 Statistical analyses

For all analyses and algorithms, two validation approaches were conducted. First, in LOSO validations, algorithms are trained on all but one participant's data and that participant is held back for validation. This process is repeated until all participants have served as the validation participant once. Second, an out-of-sample validation was performed in which the entire dataset from one study is used as training data and the second study is used for

validation. Regression algorithms were evaluated by RMSE and MAPE (See **section 3.4.5**) and concordance correlation coefficient (CCC) with the 'DescTools' package in R. Equivalence tests were employed to determine if the true METs and predicted METs were statistically equivalent and these are explained further in **section 3.5.4.** For classification tasks, the Kappa statistic, accuracy and the F1-Score are reported (see **section 3.5.4** for further details). A p-value threshold of <0.05 is used to determine statistical significance where p-values are reported.

#### 7.2.7.1 Algorithms and hyperparameter selection

The SWA outputs a METs estimate which was evaluated in this study (SWA manufacturer). Several machine learning algorithms for the regression and classification tasks were tested, which are described in **section 3.5.3**. In the regression tasks, algorithms predicted a MET value for each minute and in the classification tasks algorithms classified activity categories for each minute. The activity classifications were: Sedentary (≤1.5 METs), Light (> 1.5 and < 3 METs) and MVPA (≥3 METs) which are standard cut-offs (Beaulieu et al., 2017; Blair et al., 2014; Farrahi et al., 2020). For each algorithm, the hyperparameters were informed by a random search through a range of potential hyperparameters in preliminary tuning experiments. Random search iterates over a grid of randomly selected combinations of hyperparameters, rather than exploring every possible combination of features and therefore offers a significant computational advantage over a grid-search approach (Géron, 2019. pp 78). Each random search was conducted with the RandomizedSearchCV class in Scikit Learn (Pedregosa et al., 2011), using three-fold cross-validation. The specific parameters for each algorithm are detailed in Appendix 3.1 and except for Neural Network models (See **section 3.5.2.2**) the scoring/loss criterion was the respective loss or scoring metrics within Scikit Learn. All algorithms were trained using Keras-GPU (Chollet, 2015) or Scikit Learn (Pedregosa et al., 2011). The specific algorithms are detailed in **section 3.5.2**.

#### 7.2.7.2 Permutation importance

Permutation importance is computed for random forests using the sklearn.inspection.permutation_importance class. This method provides a means of investigating the importance of each of the predictive variables. A single random forest model is fitted for the FB dataset, the SWA dataset and the AG dataset. A baseline metric is determined for the estimator ($R^2$ of the prediction). After a feature is permuted the metric is calculated again and the outcome is the difference between these two scores.

#### 7.2.7.3 Simulation

It is important to consider the potential implications of having relative extremes in variables. For example, age, weight and height may be considered important variables that lead to variability in EE outputs. As such a simulation experiment was conducted to examine this. First, a dataset was simulated based on the activities conducted in the TEED study (i.e. it followed a similar but not identical pattern of movement and heart rate). A male and female participant was simulated with approximately average age, height and weight. This was achieved by setting the standardised score to 0 for each of these variables. Next, each of these were replaced by a value for weight, age and height which varied from –2SD to +2SD, in increments of 0.5SD. Predictions are made by the gradient boost algorithm used in chapter 8 and 9 and these predictions are saved. This process is repeated 100 times for each value and each variable, but a slightly different dataset is used, for example, a male with height -2SD from the mean will have 100 datasets generated. Variation in the datasets was achieved by adding Gaussian noise to each input variable (Mean = 0, SD = 0.3).

## 7.3 Results

### 7.3.1 Regression

A total of 89 participant activity sessions were included in this sample and all models could be evaluated on at least 5448 minutes of data in LOSO validations. The regression algorithms predicting measured EE are presented in table 7.3 and visually displayed in figure 7.1. The greatest error in METs was observed for the manufacturer provided SWA estimates; with a MAPE and RMSE of 34.54 and 1.86, respectively. For the AG, RMSE was lowest for gradient boost (0.93 METs), which also achieved the lowest MAPE of any AG model (17.88%). Of the FB models, the random forest and gradient boost had equal RMSE (1.36 METs), but a slightly lower MAPE was achieved by the random forest. For the SWA the gradient boost had the lowest RMSE value (0.91 METs) and the lowest RMSE of all those tested. In general, the neural network models were associated with greater RMSE overall for AG, FB and SWA models.

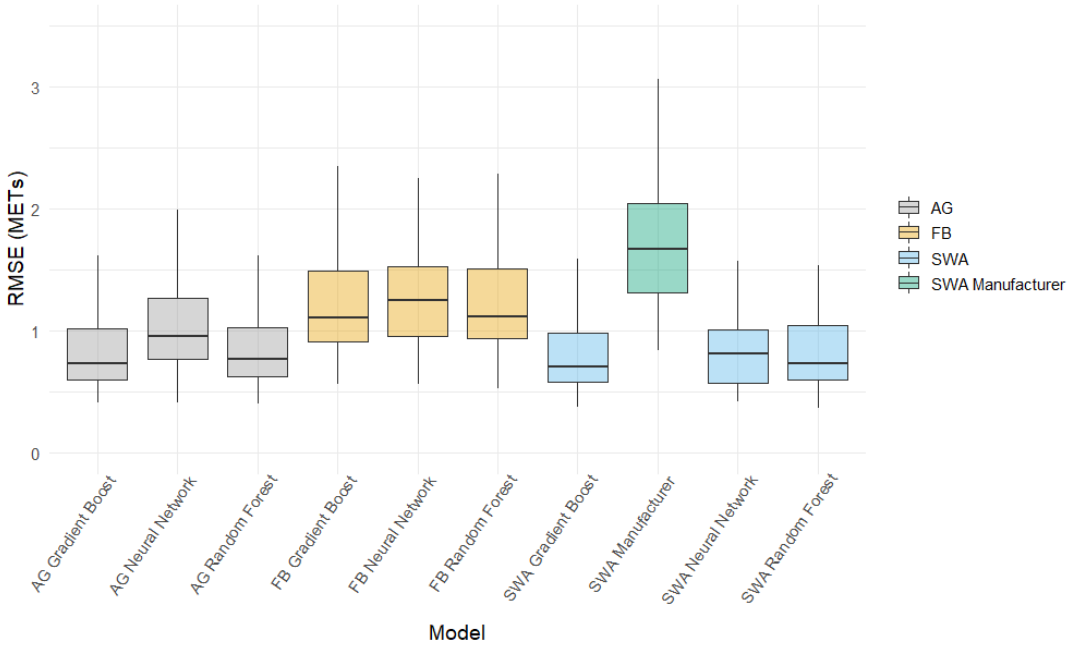



**Figure 7.1** Boxplots demonstrating the RMSE overall for each of the tested models. RMSE is calculated at the level of the subject before plotting.

Abbreviations: Root mean squared error (RMSE), Fitbit (FB), AG = ActiGraph (AG), SenseWear (SWA).

**Table 7.3** Results for each of the regression models computed across all available minutes.

Abbreviations: Fitbit (FB), ActiGraph (AG), SenseWear (SWA). Root mean squared error (RMSE), Mean absolute percentage error (MAPE), concordance correlation coefficient (CCC). Minutes refers to the number of minutes the algorithms are validated on CCC is presented with 95% confidence intervals. Equivalence implies model was statistically equivalent to the criterion.

| Model | Minutes | Participants | Predicted (METs) | True (METs) | MAPE | RMSE | CCC (95% CI) | Equivalence |
|---|---|---|---|---|---|---|---|---|
| **SWA Manufacturer** | 5533 | 88 | 3.8 ± 2.49 | 4.04 ± 2.59 | 34.54 | 1.86 | 0.73 (0.72, 0.74) | |
| **AG Gradient Boost** | 5517 | 87 | 4.04 ± 2.35 | 4.04 ± 2.59 | 17.88 | 0.93 | 0.93 (0.93, 0.93) | Equivalent |
| **AG Neural Network** | 5517 | 87 | 4.05 ± 2.55 | 4.04 ± 2.59 | 21.65 | 1.14 | 0.9 (0.9, 0.91) | Equivalent |
| **AG Random Forest** | 5517 | 87 | 4.05 ± 2.32 | 4.04 ± 2.59 | 18.36 | 0.94 | 0.93 (0.92, 0.93) | Equivalent |
| **FB Gradient Boost** | 5448 | 86 | 4.03 ± 2.19 | 4.01 ± 2.58 | 30.22 | 1.36 | 0.84 (0.83, 0.84) | Equivalent |
| **FB Neural Network** | 5448 | 86 | 4.02 ± 2.28 | 4.01 ± 2.58 | 32.27 | 1.45 | 0.82 (0.82, 0.83) | Equivalent |
| **FB Random Forest** | 5448 | 86 | 4.03 ± 2.14 | 4.01 ± 2.58 | 30.10 | 1.36 | 0.84 (0.83, 0.84) | Equivalent |
| **SWA Gradient Boost** | 5492 | 87 | 4.04 ± 2.39 | 4.04 ± 2.6 | 17.83 | 0.91 | 0.93 (0.93, 0.94) | Equivalent |
| **SWA Neural Network** | 5492 | 87 | 4.05 ± 2.47 | 4.04 ± 2.6 | 19.56 | 0.96 | 0.93 (0.92, 0.93) | Equivalent |
| **SWA Random Forest** | 5492 | 87 | 4.05 ± 2.35 | 4.04 ± 2.6 | 18.25 | 0.92 | 0.93 (0.93, 0.93) | Equivalent |

Activity-specific MET predictions are presented for minute-level data in appendix 3.2 and summarised in figure 7.2. For all activities tested, tree-based models (gradient boost or random forest) applied to AG or SWA data were superior as measured by RMSE. The manufacturer estimates of the SWA had the highest RMSE for all activities aside from sedentary activities, in which only the AG gradient boost and random forest had a lower RMSE. Notably, all FB models overestimated sedentary activities and had the highest RMSE in this category. An example of the model predictions for a single subject is shown in Figure 7.3.

Table 7.4 demonstrates the statistics for between study predictions. Notably larger errors were observed relative to the LOSO validations, with a FB model reaching a RMSE of 1.92 (Neural network) when study 1 was used as the training data.

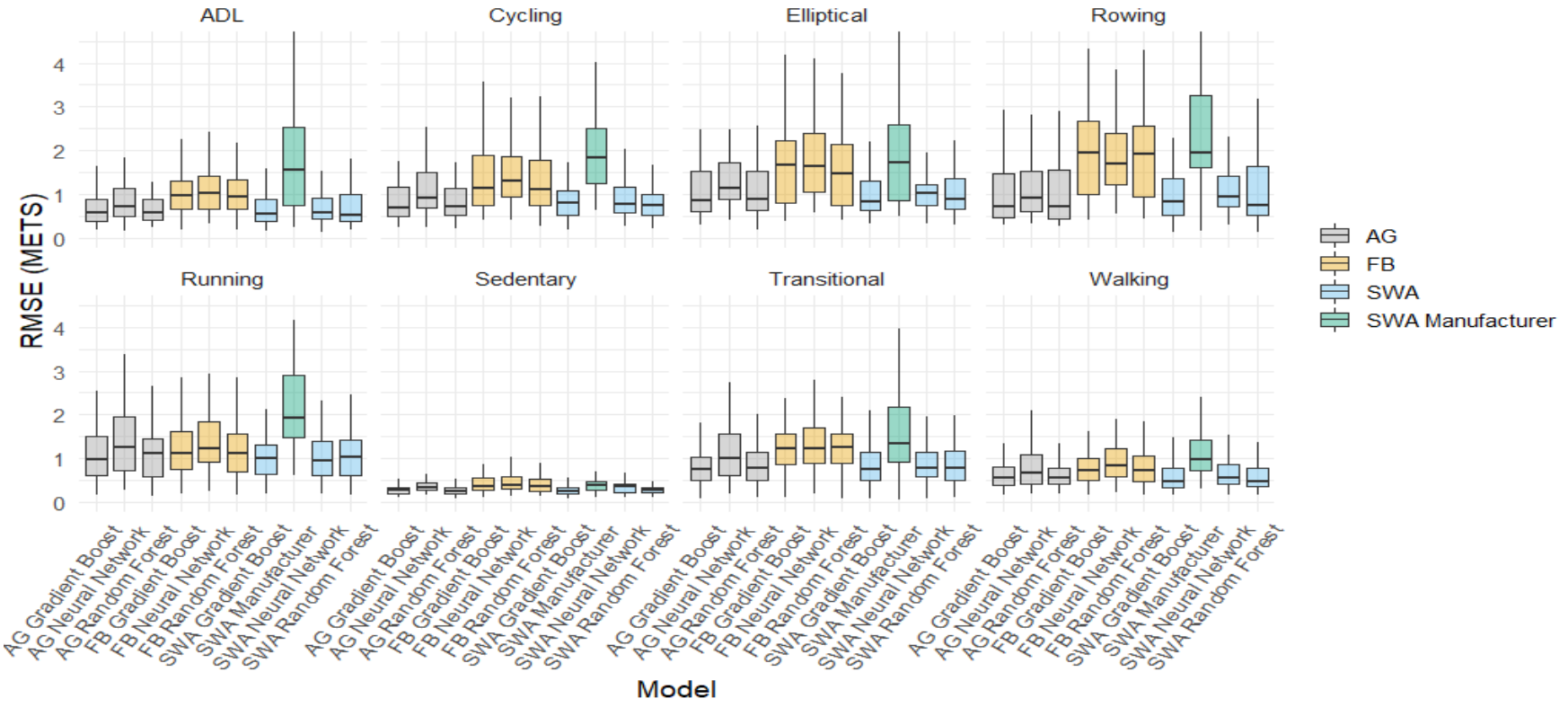



**Figure 7.2** Boxplots demonstrating the RMSE overall for each of the tested models in specific activities. RMSE is calculated at the level of the subject and activity before plotting.

Abbreviations: Root mean squared error (RMSE), Fitbit (FB), AG = ActiGraph (AG), SenseWear (SWA).

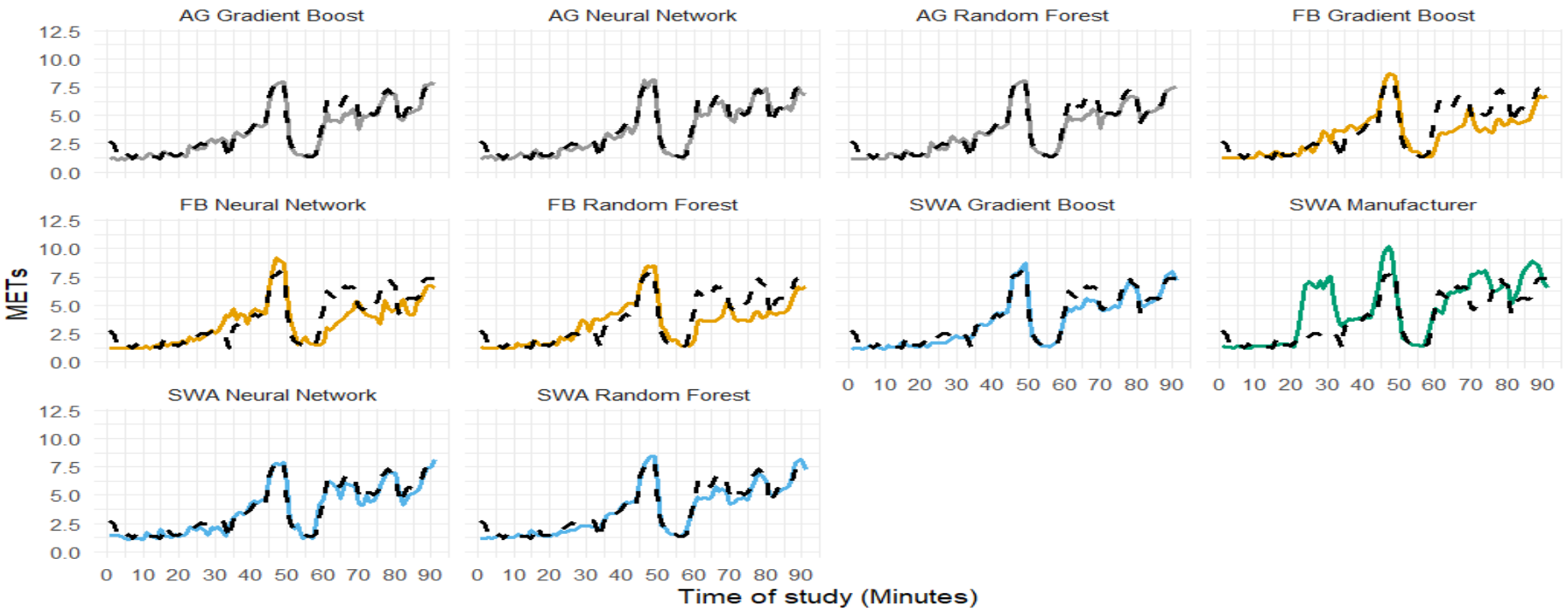



**Figure 7.3** A time series plot showing METs predicted by the models tested in this study and by indirect calorimetry (black dashed line), for a single subject in study 2.

The x-axis represents the time of measurement. Minutes 1-15 = Sedentary, minutes 16-17=transitional/unstructured, minutes 18-32 = activities of daily living (typing, wiping surfaces and ironing), minutes 33-34 = Transitional/unstructured, minutes 35-44 = Walking, minutes 45-49 = Running, minutes 50-59 = Transitional/unstructured, minutes 60-69 = Cycling, minutes 71-80 = Rowing and minutes

82-91 = Elliptical. Participants performed cycling, rowing and elliptical at self-selected low and moderate intensity for 5 minutes each.
Abbreviations: Metabolic equivalents (METs), Fitbit (FB), ActiGraph (AG), SenseWear (SWA).

### 7.3.2 Permutation importance

To estimate the relative importance of each of the features used in each model, permutation importance has been reported. The boxplots below represent the most important features according to this measure from 20 iterations and the results are shown in figures 7.4 for FB, 7.5 for SWA and 7.6 for AG. In all cases, it appears that heart rate variables are most important. A clear difference is noted between FB and AG/SWA models. In AG/SWA, heart rate is by far the most important whereas in the FB models the difference from the next best variable is small.

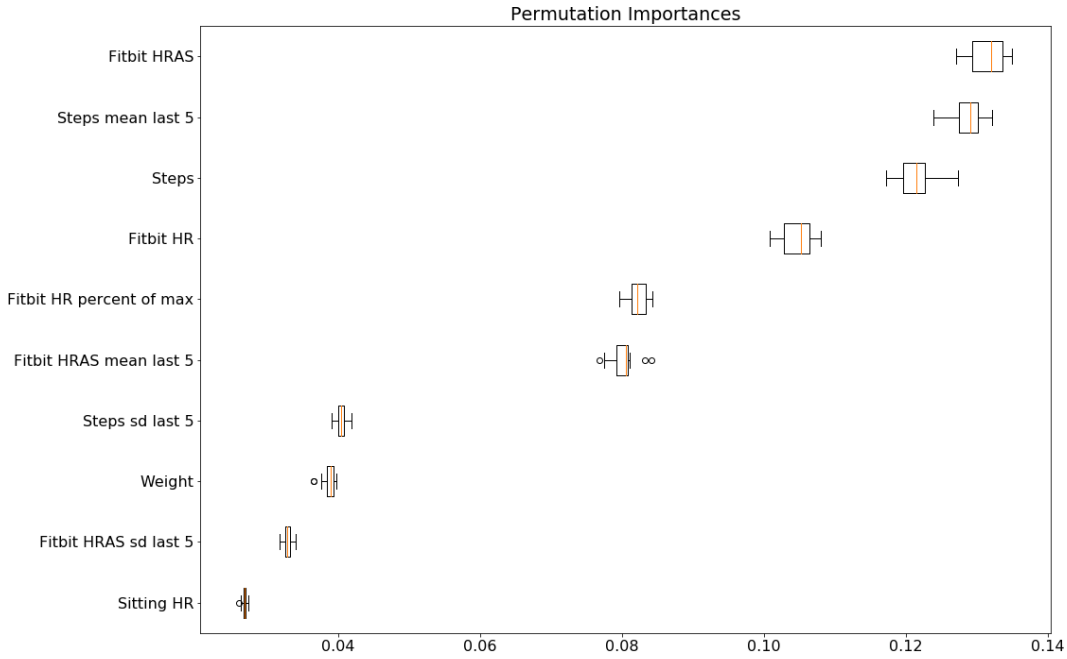



**Figure 7.4** Permutation importance for the top 10 variables in the FB dataset.

The axes represent the $R^2$ change with the permutation of the variable. Abbreviations: Heart rate above sitting (HRAS), heart rate (HR), standard deviation (SD).

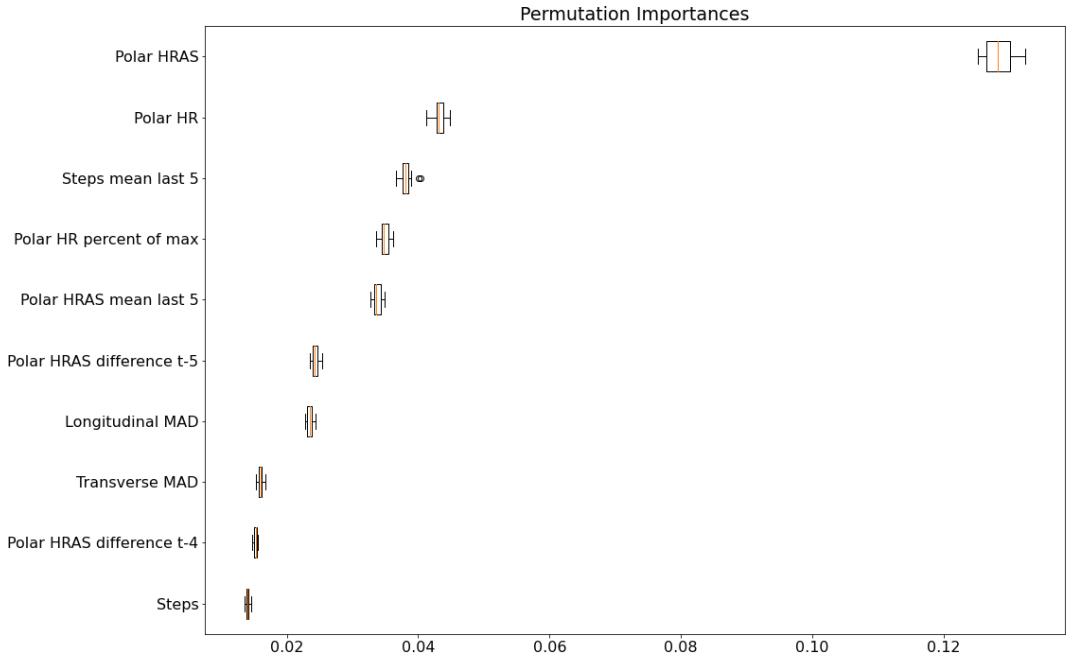



**Figure 7.5** Permutation importance for the top 10 variables in the SWA dataset.

The axes represent the $R^2$ change with the permutation of the variable. Abbreviations: Heart rate above sitting (HRAS), heart rate (HR), Mean of absolute deviation (MAD).

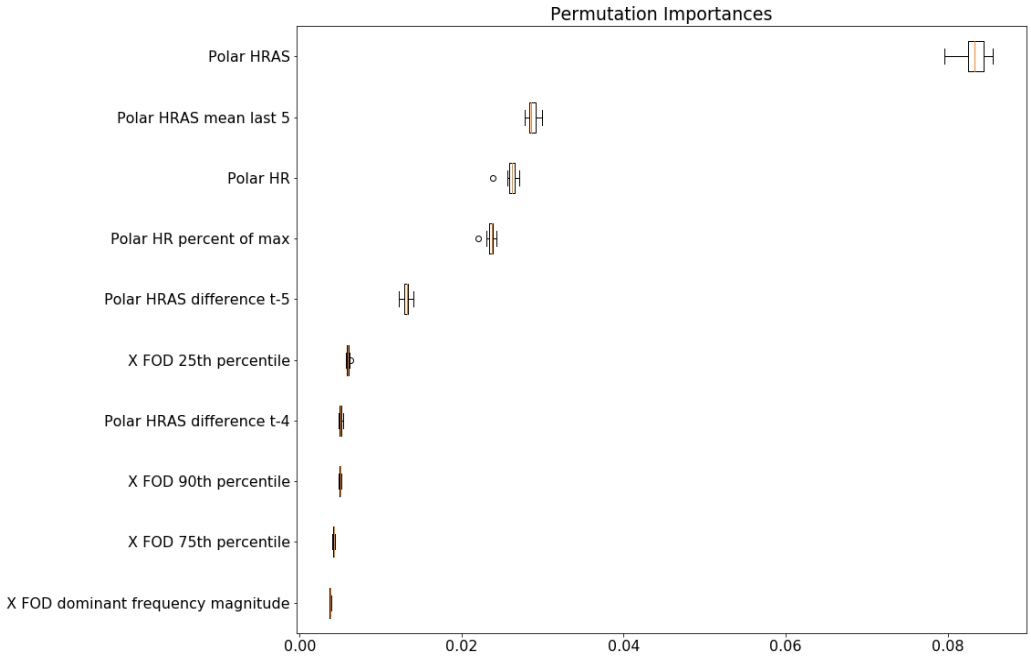



**Figure 7.6** Permutation importance for the top 10 variables in the AG dataset.

The axes represent the $R^2$ change with the permutation of the variable. Abbreviations: Heart rate above sitting (HRAS), heart rate (HR), X-Axis (X), first order differential (FOD).

### 7.3.3 Simulation of model performance

The results of the simulation are presented in figure 7.7, which represent the effect of varying static, demographic inputs and this is explained further in **section 7.2.7.3**. These plots demonstrate that the METs outputs are relatively constant between the genders, and across the range of height, weight and age. The apparent greatest difference occurs at the end of the protocol (Elliptical) for the iteration using male data and varying the weight in the input data. The mean METs here for weight -2 SD from the mean was 6.57, compared to 5.24 for 1.5 SD from the mean.

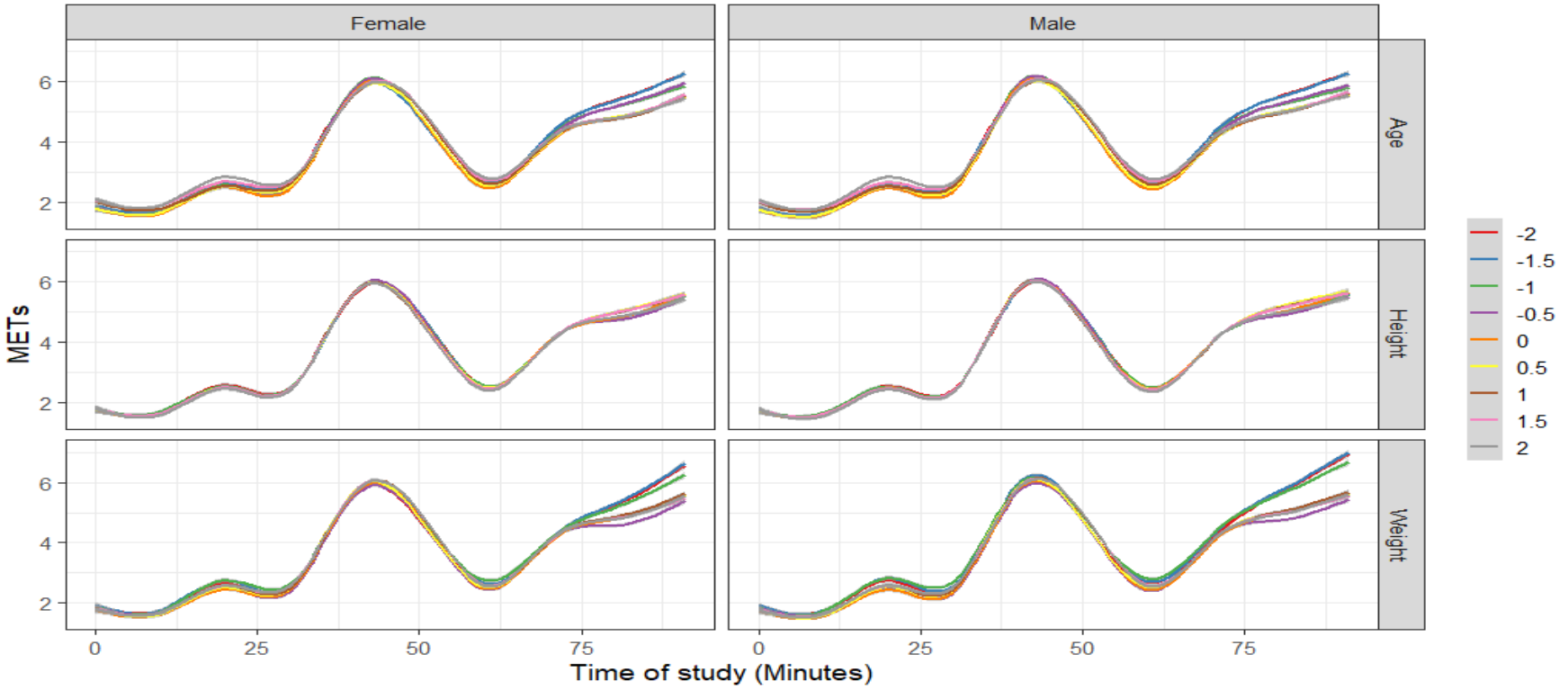



**Figure 7.7** A time series plot showing METs predicted by the Fitbit Gradient boost for a Male and Female, with varying input features. 100 simulations were conducted for each input. Line colours represent each increment in standard deviation from the mean. Lines are shown with standard error. Minutes 1-15 = Sedentary, minutes 18-32 = activities of daily living (typing, wiping surfaces and ironing), minutes 33-34 = Transitional/unstructured, minutes 35-44 = Walking, minutes 45-49 = Running, minutes 50-59 = Transitional/unstructured, minutes 60-69 = Cycling, minutes 71-80 = Rowing and minutes 82-91 = Elliptical. Participants performed cycling, rowing and elliptical at self-selected low and moderate intensity for 5 minutes each.

**Table 7.4** Out-of-sample results for each of the regression models.

Abbreviations: Fitbit (FB), ActiGraph (AG), SenseWear (SWA). Metabolic equivalents (METs), Root mean squared error (RMSE), Mean absolute percentage error (MAPE), concordance correlation coefficient (CCC). Minutes refers to the number of minutes the algorithms are validated on CCC is presented with 95% confidence intervals. Equivalence implies model was statistically equivalent to the criterion.

| Model | Training data | Minutes | Predicted (METs) | True (METs) | MAPE | RMSE | CCC (95% CI) | Equivalence |
|---|---|---|---|---|---|---|---|---|
| **AG Gradient Boost** | Study 1 | 2690 | 4.03 ± 1.9 | 3.93 ± 2.66 | 36.35 | 1.37 | 0.82 (0.81, 0.83) | Equivalent |
| **AG Neural Network** | Study 1 | 2690 | 4.07 ± 2.48 | 3.93 ± 2.66 | 29.75 | 1.33 | 0.87 (0.86, 0.88) | Equivalent |
| **AG Random Forest** | Study 1 | 2690 | 3.97 ± 1.79 | 3.93 ± 2.66 | 39.50 | 1.51 | 0.78 (0.77, 0.79) | Equivalent |
| **FB Gradient Boost** | Study 1 | 2630 | 3.76 ± 1.7 | 3.88 ± 2.65 | 47.55 | 1.89 | 0.64 (0.62, 0.66) | Equivalent |
| **FB Neural Network** | Study 1 | 2630 | 3.65 ± 1.86 | 3.88 ± 2.65 | 47.40 | 1.92 | 0.65 (0.63, 0.67) | |
| **FB Random Forest** | Study 1 | 2630 | 3.76 ± 1.66 | 3.88 ± 2.65 | 47.45 | 1.87 | 0.64 (0.63, 0.66) | Equivalent |
| **SWA Gradient Boost** | Study 1 | 2633 | 3.92 ± 2.13 | 3.94 ± 2.68 | 27.35 | 1.23 | 0.87 (0.86, 0.88) | Equivalent |
| **SWA Neural Network** | Study 1 | 2633 | 3.88 ± 2.26 | 3.94 ± 2.68 | 27.07 | 1.22 | 0.88 (0.87, 0.89) | Equivalent |
| **SWA Random Forest** | Study 1 | 2633 | 3.91 ± 2.07 | 3.94 ± 2.68 | 29.54 | 1.28 | 0.86 (0.85, 0.87) | Equivalent |
| | | | | | | | | |
| **AG Gradient Boost** | Study 2 | 2827 | 4.46 ± 2.14 | 4.15 ± 2.52 | 31.49 | 1.36 | 0.83 (0.82, 0.84) | |
| **AG Neural Network** | Study 2 | 2827 | 4.24 ± 2.56 | 4.15 ± 2.52 | 29.00 | 1.42 | 0.84 (0.83, 0.85) | Equivalent |

| Model | Training data | Minutes | Predicted (METs) | True (METs) | MAPE | RMSE | CCC (95% CI) | Equivalence |
|---|---|---|---|---|---|---|---|---|
| **AG Random Forest** | Study 2 | 2827 | 4.45 ± 2.1 | 4.15 ± 2.52 | 31.47 | 1.38 | 0.82 (0.81, 0.84) | |
| **FB Gradient Boost** | Study 2 | 2818 | 4.11 ± 2.06 | 4.13 ± 2.51 | 34.38 | 1.66 | 0.74 (0.72, 0.75) | Equivalent |
| **FB Neural Network** | Study 2 | 2818 | 4.01 ± 2.04 | 4.13 ± 2.51 | 33.10 | 1.56 | 0.77 (0.75, 0.78) | Equivalent |
| **FB Random Forest** | Study 2 | 2818 | 4.21 ± 2.04 | 4.13 ± 2.51 | 33.79 | 1.62 | 0.75 (0.73, 0.77) | Equivalent |
| **SWA Gradient Boost** | Study 2 | 2859 | 4.15 ± 2.13 | 4.14 ± 2.51 | 24.90 | 1.25 | 0.86 (0.85, 0.87) | Equivalent |
| **SWA Neural Network** | Study 2 | 2859 | 3.94 ± 2.36 | 4.14 ± 2.51 | 25.65 | 1.25 | 0.87 (0.86, 0.88) | Equivalent |
| **SWA Random Forest** | Study 2 | 2859 | 4.2 ± 2.13 | 4.14 ± 2.51 | 25.72 | 1.26 | 0.85 (0.84, 0.86) | Equivalent |

### 7.3.4 Classification

Figure 7.8 presents the results of LOSO classification experiments for all classification algorithms and the SWA manufacturer estimates. Classes were slightly imbalanced, the dataset with the most availability was comprised of 19.4% Sedentary, 22.4% Light and 58.2% MVPA and there were small differences between devices due to data availability. The highest accuracy for FB models was the random forest (78.21%), for the AG models the random forest achieved the highest accuracy (84.56%) and for SWA models, the gradient boost (85.49%) was most accurate.

Table 7.5 provides class-specific (activity intensity) statistics for each of the models. Models tended to perform worse in light activity with F1-scores ranging from 0.20 (SWA Neural Network) to 0.66 (SWA gradient boost). In sedentary activities, the F1-score was improved with a range of 0.54 (AG support vector machine) – 0.83 (four models). For MVPA, the F1-score ranged from 0.80 (AG support vector machine) - 0.93 (three models).

Between-study classification accuracy is presented in table 7.6. Generally, when study 1 served as the training data, a lower accuracy was observed. When study 1 served as the training data the accuracy ranged from 0.55 (AG support vector machine) to 0.80 (two models). When study 2 served as the training data accuracy ranged from 0.65 (AG support vector machine) to 0.79 (three models).

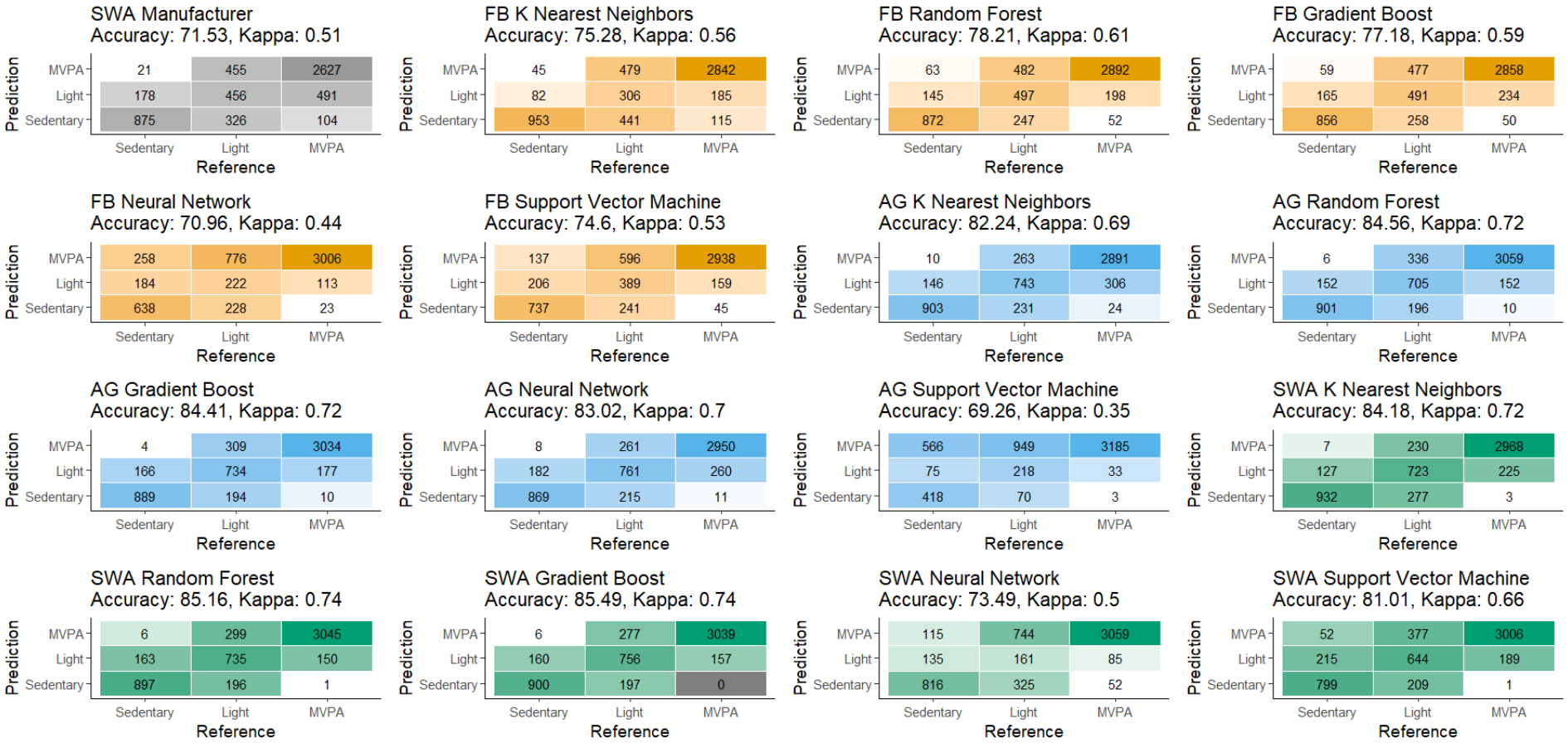



**Figure 7.8** A confusion matrix detailing the classification accuracies for each of the tested models.

Abbreviations: Fitbit (FB), ActiGraph (AG), SenseWear (SWA), Moderate to vigorous physical activity (MVPA).

**Table 7.5** LOSO results for each of the classification models.

Abbreviations: Fitbit (FB), ActiGraph (AG), SenseWear (SWA), Moderate to vigorous physical activity (MVPA).

| | Model | Sensitivity | Specificity | Precision | F1 | Balanced Accuracy |
|---|---|---|---|---|---|---|
| **Sedentary** | **AG Gradient Boost** | 0.84 | 0.95 | 0.81 | 0.83 | 0.90 |
| | **AG K Nearest Neighbors** | 0.85 | 0.94 | 0.78 | 0.81 | 0.90 |
| | **AG Neural Network** | 0.82 | 0.95 | 0.79 | 0.81 | 0.88 |
| | **AG Random Forest** | 0.85 | 0.95 | 0.81 | 0.83 | 0.90 |
| | **AG Support Vector Machine** | 0.39 | 0.98 | 0.85 | 0.54 | 0.69 |
| | **FB Gradient Boost** | 0.79 | 0.93 | 0.74 | 0.76 | 0.86 |
| | **FB K Nearest Neighbors** | 0.88 | 0.87 | 0.63 | 0.74 | 0.88 |
| | **FB Neural Network** | 0.59 | 0.94 | 0.72 | 0.65 | 0.77 |
| | **FB Random Forest** | 0.81 | 0.93 | 0.74 | 0.77 | 0.87 |
| | **FB Support Vector Machine** | 0.68 | 0.93 | 0.72 | 0.70 | 0.81 |
| | **SWA** | 0.81 | 0.90 | 0.67 | 0.74 | 0.86 |

| | Model | Sensitivity | Specificity | Precision | F1 | Balanced Accuracy |
|---|---|---|---|---|---|---|
| | **SWA Gradient Boost** | 0.84 | 0.96 | 0.82 | 0.83 | 0.90 |
| | **SWA K Nearest Neighbors** | 0.87 | 0.94 | 0.77 | 0.82 | 0.91 |
| | **SWA Neural Network** | 0.77 | 0.91 | 0.68 | 0.72 | 0.84 |
| | **SWA Random Forest** | 0.84 | 0.96 | 0.82 | 0.83 | 0.90 |
| | **SWA Support Vector Machine** | 0.75 | 0.95 | 0.79 | 0.77 | 0.85 |
| **Light** | **AG Gradient Boost** | 0.59 | 0.92 | 0.68 | 0.63 | 0.76 |
| | **AG K Nearest Neighbors** | 0.60 | 0.89 | 0.62 | 0.61 | 0.75 |
| | **AG Neural Network** | 0.62 | 0.90 | 0.63 | 0.62 | 0.76 |
| | **AG Random Forest** | 0.57 | 0.93 | 0.70 | 0.63 | 0.75 |
| | **AG Support Vector Machine** | 0.18 | 0.97 | 0.67 | 0.28 | 0.58 |
| | **FB Gradient Boost** | 0.40 | 0.91 | 0.55 | 0.46 | 0.65 |
| | **FB K Nearest** | 0.25 | 0.94 | 0.53 | 0.34 | 0.59 |

| | Model | Sensitivity | Specificity | Precision | F1 | Balanced Accuracy |
|---|---|---|---|---|---|---|
| | **Neighbors** | | | | | |
| | **FB Neural Network** | 0.18 | 0.93 | 0.43 | 0.25 | 0.56 |
| | **FB Random Forest** | 0.41 | 0.92 | 0.59 | 0.48 | 0.66 |
| | **FB Support Vector Machine** | 0.32 | 0.91 | 0.52 | 0.39 | 0.62 |
| | **SWA** | 0.37 | 0.84 | 0.41 | 0.39 | 0.61 |
| | **SWA Gradient Boost** | 0.61 | 0.93 | 0.70 | 0.66 | 0.77 |
| | **SWA K Nearest Neighbors** | 0.59 | 0.92 | 0.67 | 0.63 | 0.75 |
| | **SWA Neural Network** | 0.13 | 0.95 | 0.42 | 0.20 | 0.54 |
| | **SWA Random Forest** | 0.60 | 0.93 | 0.70 | 0.65 | 0.76 |
| | **SWA Support Vector Machine** | 0.52 | 0.91 | 0.61 | 0.57 | 0.71 |
| **MVPA** | **AG Gradient Boost** | 0.94 | 0.86 | 0.91 | 0.92 | 0.90 |
| | **AG K Nearest Neighbors** | 0.90 | 0.88 | 0.91 | 0.91 | 0.89 |

| Model | Sensitivity | Specificity | Precision | F1 | Balanced Accuracy |
|---|---|---|---|---|---|
| **AG Neural Network** | 0.92 | 0.88 | 0.92 | 0.92 | 0.90 |
| **AG Random Forest** | 0.95 | 0.85 | 0.90 | 0.92 | 0.90 |
| **AG Support Vector Machine** | 0.99 | 0.34 | 0.68 | 0.80 | 0.66 |
| **FB Gradient Boost** | 0.91 | 0.77 | 0.84 | 0.87 | 0.84 |
| **FB K Nearest Neighbors** | 0.90 | 0.77 | 0.84 | 0.87 | 0.84 |
| **FB Neural Network** | 0.96 | 0.55 | 0.74 | 0.84 | 0.75 |
| **FB Random Forest** | 0.92 | 0.76 | 0.84 | 0.88 | 0.84 |
| **FB Support Vector Machine** | 0.94 | 0.68 | 0.80 | 0.86 | 0.81 |
| **SWA** | 0.82 | 0.79 | 0.85 | 0.83 | 0.80 |
| **SWA Gradient Boost** | 0.95 | 0.88 | 0.91 | 0.93 | 0.91 |
| **SWA K Nearest Neighbors** | 0.93 | 0.90 | 0.93 | 0.93 | 0.91 |
| **SWA Neural Network** | 0.96 | 0.63 | 0.78 | 0.86 | 0.79 |
| **SWA Random** | 0.95 | 0.87 | 0.91 | 0.93 | 0.91 |

| Model | Sensitivity | Specificity | Precision | F1 | Balanced Accuracy |
|---|---|---|---|---|---|
| **Forest** | | | | | |
| **SWA Support Vector Machine** | 0.94 | 0.81 | 0.88 | 0.91 | 0.88 |

**Table 7.6** Out-of-sample results for each of the classification models. Abbreviations: Fitbit (FB), ActiGraph (AG), SenseWear (SWA).

| Training data | Model | Accuracy | Kappa |
|---|---|---|---|
| **Study 1** | AG Gradient Boost | 0.75 | 0.55 |
| | AG K Nearest Neighbors | 0.61 | 0.35 |
| | AG Neural Network | 0.72 | 0.52 |
| | AG Random Forest | 0.74 | 0.53 |
| | AG Support Vector Machine | 0.55 | 0.06 |
| | FB Gradient Boost | 0.67 | 0.43 |
| | FB K Nearest Neighbors | 0.68 | 0.47 |
| | FB Neural Network | 0.67 | 0.47 |
| | FB Random Forest | 0.67 | 0.41 |
| | FB Support Vector Machine | 0.67 | 0.45 |
| | SWA Gradient Boost | 0.80 | 0.67 |
| | SWA K Nearest Neighbors | 0.74 | 0.57 |
| | SWA Neural Network | 0.79 | 0.66 |
| | SWA Random Forest | 0.80 | 0.66 |
| | SWA Support Vector Machine | 0.68 | 0.43 |
| **Study 2** | AG Gradient Boost | 0.79 | 0.56 |
| | AG K Nearest Neighbors | 0.72 | 0.48 |
| | AG Neural Network | 0.75 | 0.51 |
| | AG Random Forest | 0.79 | 0.57 |
| | AG Support Vector Machine | 0.65 | 0.07 |
| | FB Gradient Boost | 0.73 | 0.48 |
| | FB K Nearest Neighbors | 0.72 | 0.47 |
| | FB Neural Network | 0.71 | 0.44 |
| | FB Random Forest | 0.73 | 0.48 |
| | FB Support Vector Machine | 0.73 | 0.48 |
| | SWA Gradient Boost | 0.78 | 0.57 |
| | SWA K Nearest Neighbors | 0.76 | 0.55 |
| | SWA Neural Network | 0.76 | 0.55 |
| | SWA Random Forest | 0.79 | 0.58 |
| | SWA Support Vector Machine | 0.78 | 0.55 |

## 7.4 Discussion

This study aggregated two laboratory datasets to build on previous work demonstrating the potential for machine learning algorithms to produce accurate estimates of METs and intensity class in a diverse set of activities and participants. In both regression and classification settings, the smallest errors were seen when applying tree-based algorithms (i.e. random forest and gradient boosting) to SWA and AG outputs with the RMSE and classification errors generally being higher for FB models. In almost all cases the error was smaller than the SWA manufacturer estimates. In out-of-sample generalisability experiments, greater errors and lower accuracies were observed when compared to the LOSO validations. This is the first study to classify intensity using machine learning algorithms in FB devices and accuracies up to ~78% (Kappa = 0.6) were seen in LOSO validations, with superior performance observed for sedentary and MVPA classifications. These were generally less accurate than AG and SWA models, where up to ~85% accuracy (Kappa = 0.74) was achieved. Taken together and if these results are verified in free-living studies, these findings imply that highly accurate estimates of EE, sedentary and MVPA behaviours can be estimated by all the wearables tested here.

### 7.4.1 Regression

In regression tasks, neural networks, random forests and gradient boosting were used. In previous works, neural networks and random forests have been shown to be effective in modelling EE (Ellis et al., 2014; Montoye, Begum, et al., 2017) and the present results confirm this to an extent. The RMSE values observed in the trained models ranged from 0.91 METs to 1.45 METs which improve upon the SWA manufacturer value of 1.86 METs. However, when the average METs in this study is considered (~ 4 METs), it is evident that EE prediction can be further improved. It is of note that neural networks resulted in the highest RMSE for all three devices. Similarly, Kate et al showed that neural networks resulted in bias significantly different from 0, compared with bagged decision trees and numerous other algorithms, which were not statistically different from 0 in energy cost estimation (Kate et al., 2016). Despite the utility of deep neural networks to model highly non-linear functions in some use cases, the ‘no free lunch’ theorems broadly state that there will not be an optimal algorithm for all tasks (Wolpert & Macready, 1997). Indeed, for the datasets used here, it appears tree-based ensemble models are superior for both learning tasks. It may be that the neural network models are overfitted to the training data or are being trained on insufficiently large datasets. These models are considered to have issues with overfitting in some situations,

which in some cases may be remedied by larger training sets (DeGregory et al., 2018).

Lagged accelerometer and heart rate variables were used in each model. The rate of EE depends not just on the rate of work at the point of measurement, but also on the rate of work in preceding minutes (McArdle et al., 2010, pp 172 - 234) and the relative importance of these metrics is evidenced in the variable importance analyses. Including time-lagged features allows for a clearer distinction between minutes that are relatively similar in their accelerometer pattern but differ in their measured EE, i.e. sitting for a prolonged period vs sitting after running. Transitional minutes were on average ~ 3 METs (largely attributable to the activity in the preceding minutes), compared to sedentary minutes which average ~1.3 METs, yet the error statistics were comparable to those observed in sedentary minutes, indicating that algorithms could distinguish between those minutes. This argument is further substantiated by the high placing of lagged variables in each of the permutation analyses. More advanced neural network architectures (i.e. recurrent neural networks) (Paraschiakos et al., 2020) may further the ability of models to capture the temporal dependencies of EE.

### 7.4.2 Generalisability

While many studies have reported low errors when using machine learning approaches in the estimation of EE or classification of activity, external (out-of-sample) validations are rarer and the opportunity to identify cases of overfitting has been limited. Therefore, out-of-sample validation was used between the two datasets. In all cases, degradation of performance was observed when compared to LOSO validations. Some of this reduction in accuracy is probably attributable to differences in protocols, activities and participants which means that algorithms do not have ‘similar’ minutes on which to train. Also, it is possible that the algorithms are overfitting the data. Overfitting occurs when a complex model learns the ‘noise’ in the training data that does not represent the true underlying function between inputs and the output (Vabalas et al., 2019). Previous studies have utilised out-of-sample validation or application of machine learning to accelerometer data in free-living environments (Ellis et al., 2016; Sasaki et al., 2016; Willetts et al., 2018) and errors often increase when out of sample validation is employed. Concerning the classification of physical activity intensity in multiple samples, a previous study reported reductions in out-of-sample accuracy relative to the within-sample validated models, in some algorithm and dataset comparisons (Montoye et al., 2018). However, the machine learning models still outperformed the GGIR/ENMO classification method in out of sample testing. The GGIR/ENMO method is an

established methodology based on accelerometer thresholds, which has been detailed previously (Bai et al., 2016). In another comprehensive generalisability study, 5 lab-based heterogeneous data sets were utilised to predict exercise intensity; this study found that when models were applied to a different data set than those they were generated on, model accuracy decreased from between 72-95% to between 41-60% (Farrahi et al., 2020). These drops are notably higher than in the present study and this is probably attributable to the greater differences in the accelerometer models, wear position and samples across the 5 datasets. However, caution must be exercised in a comparison between studies, as the balance of classes is likely to differ and therefore influence the evaluation metrics.

### 7.4.3 Classification

Most of the models tested in the LOSO validations show high predictive accuracy (75-85%). However, the research-grade device models (AG and SWA) were superior. FB devices provide estimates of time in each category (i.e. sedentary, light, MVPA) but the criteria and algorithms remain proprietary. Feehan et al compared estimates of time in intensities with devices such as AG, ActivPal and SWA and concluded that 80% of studies report errors of > 10% with mean differences ranging between 44% and 632% for estimations of activity above light intensity (Feehan et al., 2018). Importantly, the devices used for comparison in many studies have varying cut points and are not necessarily ‘gold-standards’ (Feehan et al., 2018). These results indicate that the application of machine learning to intensity classification can refine the large errors observed in previous studies. Despite the promising results, it must be emphasised that laboratory studies have limited ecological validity and future research should seek to address this. Whole-room indirect calorimetry would likely allow more realistic behaviours to be studied via appropriate protocols whilst providing a gold-standard comparator.

### 7.4.4 Simulation and permutation analyses

Applying these models to other datasets, where participants may have a higher or lower weight, height and age may result in substantial deviations from the expected MET output. As a MET is a standardised output relative to the subjects RMR, the MET should be relatively constant for a specific activity, indeed, the daily equivalent of this phenomena is seen in the observation that with increasing BMI, EE is increased, but as a multiple of RMR remains far more constant (Prentice et al., 1996). The standardised measures of the compendium imply that the MET output is relatively constant for a particular activity (Ainsworth et al., 2011). With this in mind, it is instructive to observe that the models produce a consistent output despite substantial variance in the subject characteristics.

The variable importance plots show the critical importance of heart rate measures, particularly for the SWA and AG models, which reflects the established relationship between heart rate and $VO_2$ (Ceesay et al., 1989). In contrast, the FB model had a smaller gap to the next most important variable. The polar heart rate strap, which was used in AG and SWA models is known to be extremely close to electrocardiogram criterion measures (Gillinov et al., 2017). Conversely, photoplethysmography-based heart rate sensors may produce “spurious” heart rate measurements (Reddy et al., 2018), which increases noise in the training and testing data sets and probably plays a significant role in reducing the importance of this variable.

### 7.4.5 Strengths

A strength of the present study is the aggregation of two data sets to provide a more comprehensive and variable data set on which to train models. Whilst the measures (sensors and indirect calorimetry) were the same between studies, the tested cohorts differed demographically and the protocols were different, which provides a good estimate of the applicability of the tested models. Combining data sets also leads to a larger number of participants (n=89), and a larger sample size than much of the previous literature (Ellis et al., 2016; Lu et al., 2018; Montoye et al., 2018; Montoye, Begum, et al., 2017; Staudenmayer et al., 2009; Zhang et al., 2012). In general, an increase in training observations is considered a mechanism of enhancing performance (Vabalas et al., 2019) and the results of the present study provide evidence that this is the case in both commercial and research-grade accelerometers.

Another strength of this study is testing numerous algorithm and device combinations. A previous study developed a multilayer neural network which was trained on a wearable system including a vest for electrocardiogram measures and 4 accelerometers (one on each wrist and thigh) (Lu et al., 2018). Despite the small bias, this is unlikely to be a feasible means of assessing free-living energy balance behaviours. Participant discomfort and sensor removal present additional biases and this may require additional modelling approaches to address (Lee, 2013; O’Driscoll, Turicchi, Duarte, et al., 2020; Xu et al., 2018). This threshold of practicality will vary depending on the size, duration, computational resource and specific aims of the research study. Therefore, the development of three models with varying requirements is a central advantage in this study.

Lastly, testing both classification and regression algorithms in the same devices enhances the use cases of the results of this study. One area of future work is to explore combined classification and regression approaches, similar to the branched

models of the Actiheart (Brage et al., 2004) or stacked ensemble approaches. This may be effective in producing refined estimates of TDEE in free-living subjects, given that most of a day is comprised of resting/sedentary minutes and some of these models slightly overestimate sedentary activities, although depending on the classification/regression methods, this could incur additional computational costs on large datasets.

### 7.4.6 Limitations

A limitation of this study was the lack of a true testing set. Rather, an estimate of the true test error is sought by i) testing on unseen participants and ii) testing on an unseen dataset. In the former, the within-subject data is far more correlated than between-subject data and this method represents the closest approximation of how such a model would perform on true test participants (Ellis et al., 2014). In the latter, this was extended so that the training set and the testing set are comprised of different participants and protocols. Beyond these validation approaches, the ultimate test of the results presented here is a free-living validation for EE and intensity class. Total daily EE can be validated with the DLW method over a 7-14 day period (Black & Cole, 2000) and the results presented here are part of a wider project discussed in the subsequent chapter. Whilst free-living validations are critical, the resolution required to evaluate activity-specific errors is obtainable from indirect calorimetry only. Regarding activity categories, no gold-standard method exists to validate time in sedentary, light and MVPA activities outside of a controlled environment and the generalisability of classification models to free-living studies is somewhat uncertain. Authors have highlighted the limitations of accelerometer data collected within a laboratory (Bastian et al., 2015; Kerr et al., 2013), the activities performed in a free-living environment are more diverse and this further necessitates the need for more naturalistic (i.e. free-living) validation studies or at least validation studies conducted over several days using diverse activity protocols in a residential facility. Lastly, to replicate predictions made by the present algorithms in free-living subjects measured RMR may be required, which increases the researcher and participant burden. A suitable alternative in the absence of measured RMR may be prediction equations derived from BMI, age, height and gender, rather than assuming a resting value of 3.5 ml $O2 \cdot kg^{-1} \cdot min^{-1}$ (Kim et al., 2017).

## 7.5 Conclusion

This study builds on previously published work to demonstrate that machine learning techniques can be used to learn the complexities of human movement and physiological data in the study of human EE. The superior performance when

datasets were combined indicates that the incorporation of additional training datasets can improve predictions. Classification and regression errors were greater when comparisons were made between studies, which may indicate a tendency to overfit a training dataset. Single-sample, cross-sectional studies generating EE models show acceptable accuracy however, it is likely that these models are overfitted to a given sample and as such, improving generalisability is essential. To extend the utility of EE estimates beyond lab conditions, more cross-testing between datasets is required, in addition to validation in free-living samples by DLW.

# Chapter 8 – Free-living validation of energy expenditure prediction models from wearable devices, a doubly labelled water study

## 8.1 Introduction

For many years, researchers have taken a wide range of approaches to model the complex phenomenon that is EE using wearable technologies. The significant milestones in this path (e.g., flex-hr, IDEEA, SWA and actigraphy models) have been outlined in **section 1.3.2** of this thesis, along with their respective limitations. In recent years, there has been a proliferation of both cloud-connected tracking devices and advanced signal processing algorithms capable of learning complex patterns in large accelerometer and physiological signal datasets (Ellis et al., 2014). Taken together, these developments could provide a method to quantify EE, and if measures of ES are available, EI in free-living humans.

The previous chapter of this thesis described the development and validation of a series of algorithms trained to predict METs, given an input of movement, demographic and physiological data. For EE prediction, accurate and precise estimates of oxygen consumption, carbon dioxide production or respiratory exchange are required for the training of algorithms, and therefore models must be trained on data collected in less ecologically valid environments (i.e., laboratories). The considerable potential of machine learning models for the estimation of EE in laboratory environments is clear (Ellis et al. 2014; O'Driscoll, Turicchi, Hopkins, et al. 2020) and this has been discussed extensively in **section 1.3.2** and shown in the models proposed in **section 7.1**. However, performance likely degrades when algorithms trained on laboratory data are applied in free-living settings, that is, the distribution of the training and testing data differ. Outside of the laboratory, behaviours are far more variable, both in terms of the type and the duration of the activity. A single minute may contain numerous activities, and behaviours may not be of a defined length as in experimental studies (Lyden et al., 2014). Furthermore, the distribution of EE in a laboratory protocol, in which participants tend to engage in moderate-to-vigorous activities, is different from what would be expected over 24 hours, where the majority of the day is spent in sleeping or sedentary activities (Matthews et al., 2008; Saidj et al., 2015). This may explain previous observations that model predictions regress towards the mean of the training data and therefore overestimate the energetic cost of resting or sedentary epochs (Montoye et al., 2015; Staudenmayer et al., 2015).

Previous studies have investigated the validity of machine learning algorithms for the estimation of EE or physical activity recognition in a simulated free-living environment. For example, Lyden et al used various machine learning algorithms in a small sample (n=13) to predict MET-hours compared to direct observation (Lyden et al., 2014). However, truly free-living studies or comparisons to the gold standard (DLW) are far rarer. In one example, White et al (White et al., 2019) investigated the validity of regression approaches to predict PAEE, comparing the estimates to DLW in a subsample from the Fenland study (Lindsay et al., 2019). The results indicated a strong agreement with a RMSE of ~ 1MJ/day for most of the models tested when resting EE estimates were included. It remains to be seen whether machine learning algorithms, which are superior to linear models in wrist-worn devices (Montoye, Begum, et al., 2017), are capable of estimating TDEE in free-living subjects.

In addition to TDEE, many health research fields are limited by the current inability to accurately derive EI in large groups of subjects (Dhurandhar et al., 2015). This is possible given two components of the energy balance equation (i.e. change in ES and TDEE). Accurate and precise estimates of ES are available through techniques such as DEXA and BodPod (see **section 1.1.3**), and therefore what is needed to 'solve' the energy balance equation is a similarly accurate estimate of TDEE. A 2018 study demonstrated that the SWA can be used to derive estimates of EI in combination with body composition data from DEXA (Shook et al., 2018). Unfortunately, the technology underlying the SWA was acquired by a competitor and production has been ceased (Welk et al., 2017). Furthermore, the methods of estimating TDEE in the SWA are completely proprietary and researchers are unclear on the methodological assumptions of this method. Thus, alternative means for estimating TDEE are required for longitudinal, large scale energy balance research.

### 8.1.1 Chapter aims

This chapter aims to evaluate several hierarchical algorithmic approaches to predict free-living TDEE in a sample of healthy adults. Subsequently, TDEE estimates are incorporated into an energy balance model as reported previously (Shook et al., 2018) to derive EI and all comparisons are compared amongst tertiles of BMI and TDEE. This chapter compares the manufacturer estimates of the FB and SWA, which also provide estimates of EE, though the methods of estimation (i.e. development cohort, development activities, algorithms) are unknown.

## 8.2 Methods

The TEED study was conducted to investigate the utility of wearable monitors coupled with statistical learning algorithms to estimate TDEE and EI. The protocol

consisted of an initial laboratory visit in which participants completed a structured exercise task consisting of a series of sedentary, light and moderate to vigorous activities. During the laboratory component EE was measured using indirect calorimetry. The data collected in this part of the study combined with a previously published laboratory study (O'Driscoll, Turicchi, Hopkins, et al. 2020) served to develop the predictive algorithms. Full details of the algorithm development and laboratory evaluation have been discussed in chapter 7. At a maximum of two months after the laboratory visit, participants returned to the laboratory where physiological measures, DLW dosing and device set up (detailed below) were completed.

### 8.2.1 Participants

The TEED study recruited 30 participants from the University of Leeds and the surrounding areas by email and word of mouth. Descriptive characteristics of the recruited sample are presented in table 8.1. This sample size was based on an equivalence test power calculation using values taken from a previous publication (Shook et al., 2018) and the alpha and beta to 5% and 20%, respectively. This test indicated that ~26 participants were required, but an additional 4 participants were recruited on the assumption that data would be lost due to device or sampling errors. Full eligibility, inclusion and recruitment procedures can be found in **chapter 3**, specifically **sections 3.1.2, 3.2.2** and **3.3.2**.

### 8.2.2 Physical measurements

All participants arrived at the laboratory for visits 2 and 3 in the fasted state and having abstained from physical activity, food intake, alcohol and caffeine for at least 12 hours. Visits 2 and 3 occurred 14 days apart. At both visits, FM and FFM in kilograms and percentage of body mass were estimated via ADP, using the BodPod and the application of the Siri model (Siri 1956). Bodyweight (± 0.1 kg) was also obtained from the BodPod scales and the change in FM and FFM over time was calculated as the difference between these two body composition measures ($\Delta\mathrm{FFM} = \mathrm{FFM}(\text{visit } 3) - \mathrm{FFM}(\text{visit } 2)$). For further details on body composition measures please see **section 3.4.1.5**. Each subject's RMR was measured on the second visit to the laboratory, with a ventilated hood indirect calorimeter system (GEM, Nutren Technology Ltd; UK) by the method described in **section 3.4.2.3**. RMR data were unavailable for one participant and in this case, RMR was approximated with BMI-specific equations (Müller et al., 2004).

### 8.2.3 Wearable devices

Participants wore several devices in the free-living component of this study. These devices are briefly explained below, but further information can be found in **section 3.5.2**. All participants were provided with a booklet describing wearing and charging instructions. Participants were instructed to remove the devices during water-based activities (i.e. showering/swimming). Participants wore a FB activity monitor securely on the non-dominant wrist. All participants were provided with a study-specific Fitbit account and were requested to synchronise their data via the mobile application daily. Minute level FB data were retrieved from the Fitbit API and stored for analysis. Participants also wore the SWA on the non-dominant upper arm and all data were downloaded upon return of the devices to the laboratory via the SWA software. Both devices were initialised with demographic information. Participants were provided with a charging cable for both the SWA and FB, should the devices run out of battery during the free-living period. If devices were removed for charging, participants were instructed to put the device back on immediately after recharging.

The Polar H10 chest strap (Polar Electro, Kempele, Finland) was used to measure heart rate continually. An ActiGraph GT9-X accelerometer (AG; ActiGraph, Pensacola, FL, USA) was worn on the non-dominant wrist, which served as a Bluetooth receiver for heart rate data and recorded acceleration at 30 Hz. Where possible, participants brought the AG and polar devices back to the laboratory for data downloading and recharging before the battery died, given the limited battery life of the AG whilst connected to a continuous heart rate monitor. Unfortunately, individual charging ports were not available for this study. Participants were asked to remove the AG and heart rate chest strap immediately before bed and put it on immediately upon waking, this places the device in ‘idle’ mode and facilitates a longer data collection period without a recharge. Upon return of devices to the laboratory data were retrieved and were downloaded for analysis via the Actilife software.

### 8.2.4 Data requirements, inclusion and imputation

Given that some participants removed devices overnight for comfort reasons, inclusion criteria were set based on availability in waking hours, which was defined as 08:00 am – 22:00 pm. A maximum of 3 hours of missing data was permitted during waking hours and any day in which this criterion was not met was not included in the analyses. This is a more stringent criterion than previous studies, (i.e. > 10 hours of waking time (Ostendorf et al., 2019)). For device/model TDEE averages to be included in this analysis participants must have provided at least 4 days of valid data including at least one weekend day. For considerations relating to

data availability please see **chapter 6**. Participants were provided with an activity log, in which they entered details about device removal (i.e. date, time, activity descriptions and duration) if they had to remove the device for any reason. The number of observed days for each of the models is reported in the results of this chapter. Where no signal was obtained from the devices (implying non-wear time) and the participant recorded activity in the missing data log, the activities were coded according to nearest matches in the compendium of physical activities (Ainsworth et al., 2011). The MET values obtained were multiplied by RMR to provide estimates of caloric expenditure in the period of removal and the non-resting component of this value was appended to the participant data, as RMR had already been used to fill missing periods (see below for justification). Where the logged missingness overlapped with observed data (e.g. a participant reported that they were not wearing the device when the sensors reported that they were), the observed device data was utilised rather than the diary estimates. The number of logged minutes used in the analysis is reported in the results of this chapter. Small missing periods occurring in heart rate data used by the models were filled by linear interpolation for instances where the gap was <10 minutes, assuming that this represented loss of contact to the wrist. Imputation strategies do exist to address missingness (Maeda et al. 2019; O'Driscoll, Turicchi, Duarte, et al. 2020) although a conscious decision was made in this chapter to fill missing gaps with resting values (i.e., RMR per minute), rather than utilising imputation algorithms. The reason for this is that the SWA and FB manufacturer estimates are by default filled with resting values and this minimises the potential for imputation strategies to influence the TDEE results, rather than actual estimates of EE.

### 8.2.5 Prediction settings

The algorithms presented in this chapter can be considered to be hierarchical approaches to making predictions on multivariate time-series data, as the method of prediction depends on a prior classification. For each minute in the dataset, many inputs were available including subject characteristics, acceleration, physiological and time-lagged features and the goal was to estimate a MET value, which can subsequently be converted to kilocalories and summed to provide TDEE. In the development studies described in **chapter 7**, it was evident that most EE models result in overestimates of EE for the most sedentary activities (i.e. <1.5 METs) (O'Driscoll, Turicchi, Hopkins, et al. 2020). Given that a substantial amount of time can be spent in sedentary behaviours (Jefferis et al., 2015) this is likely to result in overestimates of true TDEE. With this in mind and the knowledge that the variation in EE with resting behaviours is much lower than activity behaviours, a k-nearest neighbors (k-NN) classifier was used in each model to classify activities as

sedentary, light or moderate-to-vigorous activity. The details of the respective k-NN classifiers are outlined in **section 3.5.2** and **section 7.2.7.1**. Minutes classified as sedentary are assigned a MET value from a distribution which was weighted such that 1000 draws yielded a mean (SD) of 1.157 (0.123) METs. The rationale for this parameter is that sedentary minutes are most likely to be sitting (closer to 1 MET) than standing (closer to 1.5 METs) as suggested by a recent study observing healthy European office workers (Johansson et al., 2020). Minutes, where the heart rate was below the sitting heart rate, were assigned MET values from a slightly different distribution, 1000 draws of which gave a mean (SD) of 1.056 (0.113) METs. This distribution was parameterised like this because the reduced heart rate compared to the sitting heart rate implies that the subject was engaged in prolonged sedentariness or lying flat (Jones et al., 2003) and this is associated with a reduced energetic cost (closer to RMR), compared to sitting (de Almeida Mendes et al., 2018). The third distribution used for METs estimates was used to estimate sleeping minutes. Sleeping minutes also represent a deviation from the training data and were identified differently depending on the device. For FB models, minutes of sleep was determined by the FB sleeping algorithm, from sleep start time to sleep end time. In using the FB's sleep algorithm for FB models only the FB models can be considered to be 'stand-alone', so they do not require additional sensor inputs. For SWA and AG models, a different method to identify sleep was used. First, a period of sleep was identified from the last observation of heart rate at night between 21:00 and 5:00, where the period of removal was greater than 120 minutes, to the first observation of heart rate from the polar device the next morning, as participants were required to put them on immediately upon waking. As some participants wore devices all night, the SWA's sleeping algorithm was used to identify sleeping periods. If the SWA algorithm reported that the subject was sleeping but the polar heart rate was worn (participants were instructed to remove this when sleeping), this was considered to be sleep. Given the reduced metabolic rate during sleep (Goldberg et al., 1988) the distribution of METs added had a mean (SD) of 0.958 (0.11) based on 1000 draws. To each of these distributions, Gaussian noise (Mean=0, SD=0.1) was added to simulate natural variation in minute by minute EE. The data collected for the algorithm development showed variance in EE measured by indirect calorimetry within-subjects at the minute level and adding noise to the expected mean values simulates these occurrences. The resulting distributions are shown as histograms in appendix 4.1. Minutes classified as light or MVPA (i.e. > 1.5 METs) were predicted using the regression algorithms, this method is summarised and described in figure 8.1.

#### 8.2.5.1 Maximum heart rate

To estimate maximum heart rate, the Tanaka method was used (heart rate maximum = 208 - 0.7 x age) (Tanaka et al., 2001). However, in some instances, participants were observed to have higher heart rates than the equation predicted. In this case, the maximum heart rate was considered to be the mean of the 5 highest observations observed in the 14-day free-living period. This process was conducted for the FB heart rate (used in FB models) and separately for the polar heart rate (used in SWA and AG models). Again, a distinction is made between the feature inputs for the SWA/AG models and the FB models, where the polar heart rate is used for all heart rate variables in the SWA/AG models and the FB heart rate is used for FB models. This ensures that the FB models are 'stand-alone' models, requiring no external sensor inputs.

#### 8.2.5.2 Sitting heart rate

The models used in this study are calibrated relative to the sitting heart rate. This was motivated by the established relationship between the sitting heart rate and the flex point (Rennie et al., 2001). Indeed, the importance of this is shown in the variable importance plots in chapter 7 (Figures 7.4 – 7.6), where the heart rate above sitting heart rate variables show higher permutation importance than absolute heart rate. Heart rate and cardiac function can vary with alcohol consumption, environmental factors and stress (Lee et al., 2014; Ryan & Howes, 2002; Schnell et al., 2013) and therefore required updating throughout the observation period. A time series was initialised with the laboratory-measured sitting heart rate. Next, sitting heart rate on each day was estimated by taking non-moving, non-sleeping minutes and averaging those heart rates for the Fitbit and polar independently. The data were subsequently smoothed by fitting a locally weighted smoothing regression model using a fraction value of 30% of the data.

#### 8.2.5.3 Classification and regression

The last step in the classification of all minutes was to use a k-NN classifier, which was specific to each device. The purpose of this algorithm was to segment the dataset into sedentary or non-sedentary minutes and details of the k-NN classifiers have been outlined in chapter 7. After this classification, three machine learning models per device were used to predict the MET values on the 'prediction' minutes. Specifically, prediction minutes refers to non-sedentary, non-sleeping minutes where sensor data was available. The algorithms employed were artificial neural networks, gradient boosting regression and random forest regression and each of these models were tested per device, for a total of 9 algorithms. The hyperparameters and features used are the same as the leave one subject out cross-validations in chapter

7 and a flowchart of the computation approach used in this study is shown in figure 8.1.

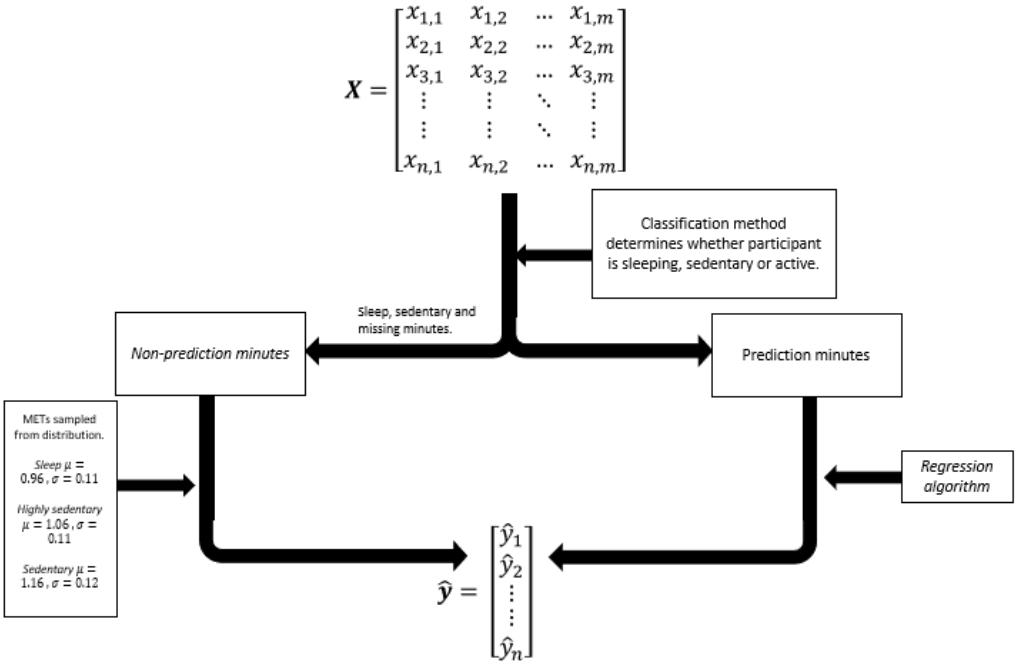



**Figure 8.1** A flowchart demonstrating the derivation of METs predictions.

The input matrix $\mathbf{X}$ contains $n$ time points and $m$ predictive features, which varies by device. Minutes are defined as 'prediction minutes' or 'non-prediction minutes' by the classification algorithm. Non-prediction minutes are sampled from a relevant distribution and this is determined by the sleeping detection algorithms or the K-NN classifiers. The specific distributions are i) Sleeping, ii) Highly sedentary (heart rate below sitting heart rate), iii) Sedentary. If a subject reports removal of the devices for a specific reason (i.e. showering), the algorithm will replace missing minutes with the relevant MET value obtained from the compendium of physical activities. The regression algorithms (Neural networks, Gradient boost, Random forest) are applied to all other minutes. The time series are subsequently concatenated to provide a vector of METs estimates $\hat{\mathbf{y}}$.

#### 8.2.5.4 Derivation of kilocalories and physical activity level

Metabolic equivalents were converted into minute-level caloric expenditure by multiplying by the RMR value per minute. All minutes were subsequently summed to provide a daily TDEE and divided by 0.9, as a rough approximation of the additional energy costs of digestion, digestive and biological processes associated with food

intake (see justification below). In addition to evaluating the TDEE estimates, the PAL is also reported, where PAL = TDEEmodel/RMRmeasured.

### 8.2.6 Dietary induced thermogenesis

Dietary induced thermogenesis, which is discussed in **section 1.1.2.3**, represents 5-15% of TDEE, assuming a mixed diet and is often approximated at 10% in energy balance research (Westerterp et al. 2004). Based on this, 10% adjustments were made to the TDEE values predicted by the models. The rationale for this is considered below.

Previous studies have taken different approaches to this issue depending on the calibration data, for instance: The Actiheart calibration data was collected when the subjects were in the fed state. One study states that: *'According to guidance provided by the manufacturer, calibration was conducted with subjects in the fed state to provide estimates of AEE with DIT included (T. Evans, CamNtech Ltd, personal communication)'* (Löf et al., 2013), and therefore DIT adjustment is not made in some instances (Löf et al., 2013). By contrast, Whybrow and colleagues add 10% of EI to the IDEEA system's estimate of TDEE as it does not account for TDEE (Whybrow, Ritz, Horgan, & Stubbs, 2013) and White and colleagues adjust their estimates to account for DIT (White et al., 2019).

The calibration data and the development of the models used in this study have been described in **chapter 7**. The two studies contributing the calibration datasets have been described in the 'TEED study' and 'Device validation study' paragraphs between **sections 3.1 – 3.3** and in **section 7.2.1**. In the first study (Device validation study, n=59) all participants arrived having fasted for at least 12 hours, then body composition and RMR measures were conducted before the exercise test. In the TEED study, participants were required to attend the laboratory after fasting for at least 4 hours or after an overnight fast where possible, which surpasses the 3 hours recommended for exercise testing (Fletcher et al., 2001). It is important to contemplate the extent to which this may be influencing the results of this study. Based on an analysis of 131 test meals of varying composition, Reed and Hill conclude that the effects of DIT may last beyond 5 hours and provide an equation describing the EE above RMR attributable to DIT: $EE\ above\ RMR\ (Kj/h) = 175.9 \times t \times e^{-t/1.3}$ (Reed & Hill, 1996). Figure 8.2 shows this model converted to kcal/min. Based on this, consider a subject with an RMR of ~2000kcal, (based on the mean weight and age of 88.6kg and 38.1 years reported in this study (Reed & Hill, 1996)), who had eaten exactly 4 hours before arrival. At 4.5 hours, if this subject was standing at an EE of 1.35 METs, the effect of DIT would increase the METs estimate by 5.27%. As EE increases (e.g. 4 METs) percentage error would decrease to

~1.78% and at 8 METs, the error would be ~0.89%. It is important to state that this represents a 'worst-case scenario' and most of the participants included in this training data would be unaffected. The average estimate is likely to be far closer to the estimate of 10%. Nonetheless, an analysis has been reported in this chapter which presents the results across a range of DIT from 0-10% of TDEE.

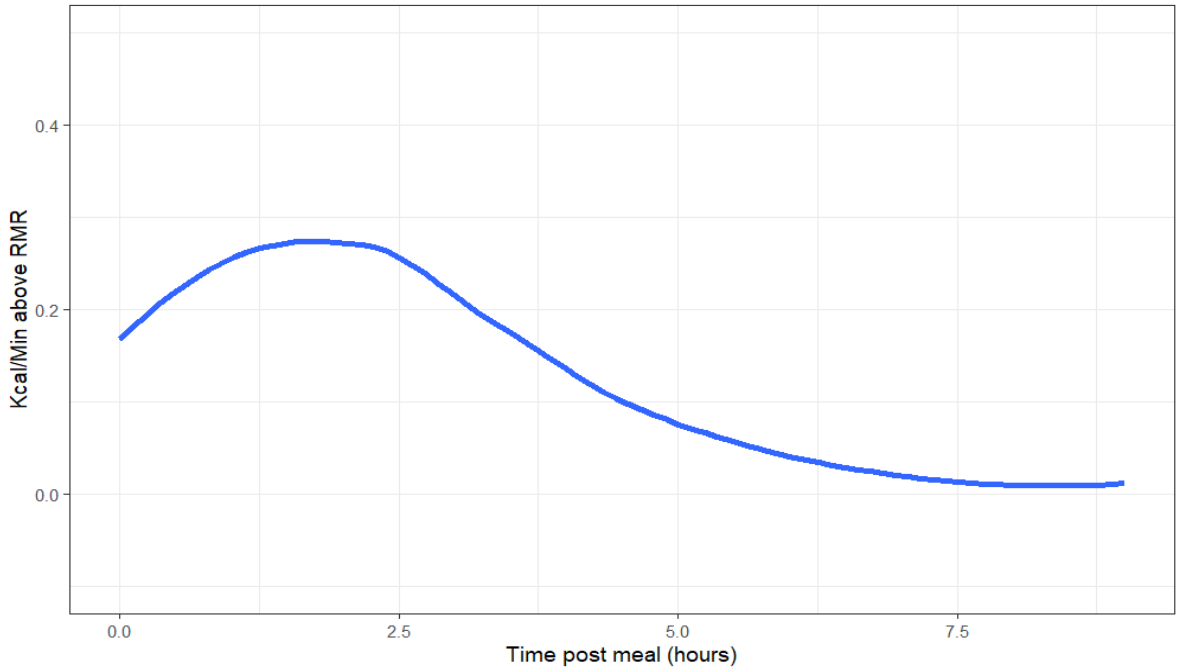



**Figure 8.2** A plot demonstrating the kcal/min above RMR after the consumption of food.

The model used for this plot is $EE\ above\ RMR\ (Kj/h) = 175.9 \times t \times e^{-t/1.3}$ estimates are then converted to kcal and divided by 60 to give kcal/min (Reed & Hill, 1996).

### 8.2.7 Energy expenditure with the SWA rather than doubly labelled water

The DLW method, as described in **chapter 3,** was intended to be the criterion measure of TDEE in this study. Unfortunately, the COVID-19 pandemic has delayed the analysis of the DLW samples and at the time of writing, DLW results to validate TDEE, PAL and EI estimates against are unavailable. This chapter, therefore, takes a comparative approach between models and to the SWA. In some populations, the SWA has been demonstrated to provide accurate estimates of TDEE (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020). However, some evidence suggests that the accuracy of the SWA may differ depending on the activity status of the individual (with limitations in highly active subjects and high-intensity activity), and a tendency towards underestimation with increasing TDEE (Koehler et al., 2015; Shook et al.,

2018). Thus, the SWA is used as a comparator in this chapter, rather than a criterion.

### 8.2.8 Energy intake

Using all TDEE estimates, the energy balance principle was applied to approximate the EI of each subject during the 14-day free-living study by the following equation (Shook et al., 2018).

$$EI = 1020 \frac{\Delta FFM}{\Delta t} + 9500 \frac{\Delta FM}{\Delta t} + EE$$

Where $\Delta\mathrm{FFM}$ and $\Delta\mathrm{FM}$ represent changes to body composition (kg), between the start and end of the assessment period, 1020 (kcal/kg) and 9500 (kcal/kg) are the assumed energy densities of FFM and FM, respectively (Thomas et al., 2010). The parameter $\Delta t$ represents days between body composition measurements. The EE term is derived from each of the algorithms as well as the SWA and FB manufacturer estimates. This analysis is similar to the TDEE analysis explained above, however, it is included to align with previous studies (Shook et al., 2018) and because a change in weight or body composition (i.e. $\frac{\Delta FFM}{\Delta t} \neq 0$ or $\frac{\Delta FM}{\Delta t} \neq 0$) would imply tissue gain or loss and therefore EI ≠ EE.

### 8.2.9 Statistical analyses

Unless otherwise stated, data are presented as means ± sd. Agreement between measures was assessed by the method of Bland and Altman (Altman & Bland, 1983) relative to the SWA. Statistical equivalence tests with a ± 10 % equivalence bound were used to determine if methods were equivalent. The metrics RMSE, MAPE and Pearson correlations were also used to evaluate the agreement. Patterns in EE predictions were compared between each of the models across different levels of activity. These cut-offs were determined by the SWA METs estimate and the cut-offs were: Sedentary ≤1.5 METs, Light >1.5 METs and <3 METs and MVPA ≥ 3 METs, for this analysis, the average EE estimate for each model in each MET cut-off was used.

Variable importance plots in **chapter 7** confirmed that heart rate is an important predictor of EE. Indeed, the relationship shows extremely close associations during activity and has been recognised for many decades (Leonard, 2003). Thus, estimates of EE should be highly correlated with the subject's measured heart rate. To investigate this, within-subject Pearson's correlations for the relationship between polar heart rate and EE predictions are reported for all minutes classified as MVPA.

To compare models at different levels of TDEE (estimated by SWA) and BMI, the sample was tertiled into approximately equally sized groups as has been done previously (Shook et al., 2018). Pre-processing and application of algorithms were conducted in Python 3.7.6, using the Keras-GPU (Chollet, 2015) library for neural networks or Scikit Learn (Pedregosa et al., 2011) for other machine learning algorithms. Statistical analysis and visualisations were conducted in R version 3.6.3. A p-value of $<0.05$ is used to determine statistical significance where p-values are reported.

## 8.3 Results

### 8.3.1 Sample

The descriptive characteristics of the whole sample, split by gender, and as tertiles of BMI and TDEE, are presented in table 8.1 and the sample and tertiled averages for TDEE are presented in table 8.2. Based on the inclusion criteria stated above, 28 participants could be included in the AG and SWA machine learning models analysis, 29 participants were available for the SWA manufacturer, 30 for the FB machine learning models and 30 for the FB manufacturer. Generally, participants were weight stable over the 14-day measurement period with a mean change in weight of +0.3 ± 1.1 kg. Participants in all groups averaged above 10,000 steps/day (as measured by the SWA, calculated on included days only and all participants were included), however, a wide range was observed in average steps/day from 2580 steps/day to 21798 steps/day for individual participants.

**Table 8.1** Descriptive characteristics of the included sample.

Weight refers to the weight recorded at visit 2 of the teed study. Abbreviations: Total daily energy expenditure (TDEE), Body mass index (BMI), Fat-free mass (FFM), Fat mass (FM), Resting metabolic rate (RMR), Male (M), Female (F). TDEE high has n=9 as one participant did not meet the inclusion criteria for the SWA.

| Group | N (female) | Age | Height | Weight | BMI ($kg/m^2$) | FFM (kg) | FM (kg) | FM (%) | RMR (kcal/day) | Steps/day |
|---|---|---|---|---|---|---|---|---|---|---|
| **All** | 30 (13) | 31.87 ± 10.23 | 171.86 ± 9.21 | 70.15 ± 12.88 | 23.68 ± 3.59 | 55.01 ± 12.56 | 15.14 ± 7.1 | 21.74 ± 8.73 | 1769.29 ± 435.82 | 12030.27 ± 4635.18 |
| **F** | 13 (13) | 33.31 ± 10.65 | 164.52 ± 5.29 | 60.97 ± 11.01 | 22.6 ± 4.46 | 43.78 ± 5.65 | 17.19 ± 7.86 | 27.35 ± 7.41 | 1444.69 ± 238.45 | 13446.93 ± 5933.45 |
| **M** | 17 (0) | 30.76 ± 10.09 | 177.48 ± 7.45 | 77.17 ± 9.42 | 24.51 ± 2.61 | 63.6 ± 8.99 | 13.57 ± 6.25 | 17.45 ± 7.19 | 2017.51 ± 387.99 | 10946.95 ± 3107.96 |
| **TDEE low** | 10 (9) | 31.8 ± 8.78 | 163.96 ± 3.15 | 60.52 ± 11.58 | 22.59 ± 4.8 | 44.47 ± 5.56 | 16.05 ± 9.07 | 25.48 ± 8.57 | 1522.36 ± 187.84 | 11801.37 ± 6433.39 |
| **TDEE med** | 10 (3) | 35.5 ± 12.71 | 173.92 ± 8.92 | 70.15 ± 9.68 | 23.25 ± 3.16 | 55.51 ± 8.59 | 14.64 ± 5.74 | 20.8 ± 7.09 | 1699.46 ± 399.05 | 12019.27 ± 4122 |
| **TDEE high** | 9 (0) | 28.67 ± 8.79 | 179.56 ± 6.51 | 81.92 ± 7.82 | 25.42 ± 2.12 | 68.03 ± 9.43 | 13.89 ± 6.64 | 16.98 ± 8.1 | 2197.94 ± 347.97 | 12245.39 ± 3472.94 |
| **BMI low** | 10 (7) | 27 ± 4.32 | 171.5 ± 9.71 | 58.87 ± 9.16 | 19.93 ± 1.7 | 46.91 ± 10.57 | 11.96 ± 3.16 | 20.85 ± 6.16 | 1479.42 ± 387.77 | 13643.72 ± 4959.16 |
| **BMI med** | 10 (4) | 30.1 ± 7.59 | 170.56 ± 8.13 | 69.65 ± 7.14 | 23.89 ± 0.6 | 55.59 ± 10.25 | 14.07 ± 5.13 | 20.64 ± 8.65 | 1787.39 ± 351.3 | 11359.75 ± 3771.79 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **BMI high** | 10 (2) | 38.5 ± 13.51 | 173.52 ± 10.39 | 81.92 ± 10.19 | 27.23 ± 2.94 | 62.54 ± 12.44 | 19.38 ± 9.68 | 23.73 ± 11.23 | 2041.05 ± 403.06 | 11087.35 ± 5101.9 |

**Table 8.2** Total daily energy expenditure (TDEE) estimates for each model included in this study

Results are presented as kcal/day for low, medium and high tertiles for TDEE and BMI cut-offs. Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA), Total daily energy expenditure (TDEE), Body mass index (BMI). N describes the number of participants used to derive means and standard deviations.

| | **AG Gradient Boost** | **AG Neural Network** | **AG Random Forest** | **FB Gradient Boost** | **FB Neural Network** | **FB Random Forest** | **FB Manufacturer** | **SWA Gradient Boost** | **SWA Neural Network** | **SWA Random Forest** | **SWA Manufacturer** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **All** | 3291.14 ± 787.21 , n= 28 | 3303.4 ± 811.52 , n= 28 | 3288.51 ± 774.87 , n= 28 | 3038.06 ± 723.18 , n= 30 | 2993.04 ± 736.99 , n= 30 | 3058.11 ± 740.77 , n= 30 | 2781.48 ± 608.61 , n= 30 | 3089.7 ± 698.41 , n= 28 | 3074.88 ± 694.06 , n= 28 | 3104.76 ± 697.89 , n= 28 | 2927.94 ± 492.38 , n= 29 |
| **TDEE low** | 2837.32 ± 432.53 , n= 10 | 2886.28 ± 548.96 , n= 10 | 2861.02 ± 436.98 , n= 10 | 2647.98 ± 407.97 , n= 10 | 2589 ± 421.48 , n= 10 | 2668.83 ± 418.97 , n= 10 | 2218.59 ± 230.35 , n= 10 | 2711.42 ± 451.43 , n= 10 | 2750.46 ± 510.01 , n= 10 | 2742.96 ± 458.84 , n= 10 | 2400.31 ± 225.54 , n= 10 |
| **TDEE med** | 3083.61 ± 559.84 , n= 10 | 3046.18 ± 546.81 , n= 10 | 3077.4 ± 590.36 , n= 10 | 2860.45 ± 539.92 , n= 10 | 2845.55 ± 551.96 , n= 10 | 2883.95 ± 575.36 , n= 10 | 2808.99 ± 337.47 , n= 10 | 2883.85 ± 515.12 , n= 10 | 2835.29 ± 499.97 , n= 10 | 2892.88 ± 539.63 , n= 10 | 2971.89 ± 155.27 , n= 10 |

| | AG Gradient Boost | AG Neural Network | AG Random Forest | FB Gradient Boost | FB Neural Network | FB Random Forest | FB Manufacturer | SWA Gradient Boost | SWA Neural Network | SWA Random Forest | SWA Manufacturer |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **TDEE high** | 4117.81 ± 781.04 , n= 8 | 4146.32 ± 783.07 , n= 8 | 4086.76 ± 754.54 , n= 8 | 3765.05 ± 680.19 , n= 9 | 3720.56 ± 682.9 , n= 9 | 3790.22 ± 689.26 , n= 9 | 3432.5 ± 499.16 , n= 9 | 3819.87 ± 635.61 , n= 8 | 3779.89 ± 633.89 , n= 8 | 3821.87 ± 628.23 , n= 8 | 3465.35 ± 298.1 , n= 9 |
| **BMI low** | 2833.14 ± 689.2 , n= 10 | 2900.02 ± 742.44 , n= 10 | 2827.26 ± 695.67 , n= 10 | 2662 ± 582.24 , n= 10 | 2663.46 ± 615.46 , n= 10 | 2666.42 ± 601.65 , n= 10 | 2405.51 ± 399.35 , n= 10 | 2680.67 ± 675.96 , n= 10 | 2697.13 ± 662.57 , n= 10 | 2691.49 ± 687.47 , n= 10 | 2657.34 ± 399.25 , n= 10 |
| **BMI med** | 3511.81 ± 722.38 , n= 9 | 3567.61 ± 779.69 , n= 9 | 3514.3 ± 683.93 , n= 9 | 3029.19 ± 563.19 , n= 10 | 2973.52 ± 626.66 , n= 10 | 3055.76 ± 576.69 , n= 10 | 2870.17 ± 614.59 , n= 10 | 3284.25 ± 599.68 , n= 9 | 3318.5 ± 623.07 , n= 9 | 3301.31 ± 583.27 , n= 9 | 3013.33 ± 526.05 , n= 9 |
| **BMI high** | 3579.36 ± 790.47 , n= 9 | 3487.39 ± 821.34 , n= 9 | 3575.22 ± 773.59 , n= 9 | 3422.99 ± 840.93 , n= 10 | 3342.12 ± 849.79 , n= 10 | 3452.15 ± 854.77 , n= 10 | 3068.74 ± 631.54 , n= 10 | 3349.64 ± 666.98 , n= 9 | 3250.99 ± 680.93 , n= 9 | 3367.4 ± 661.3 , n= 9 | 3121.68 ± 467.45 , n= 10 |

### 8.3.2 Data availability

The data available for modelling differs between the sensors used. Of the included participants, AG models had a mean of 8.9 ± 2.1 days, FB models had a mean of 12.4 ± 2.1 days, SWA models had a mean of 8.6 ± 2 days, SWA manufacturer estimates had 10.8 ± 1.5 days and FB manufacturer had 11.9 ± 2.1 days. To display days of data available per subject, an empirical cumulative distribution plot is shown in figure 8.3.

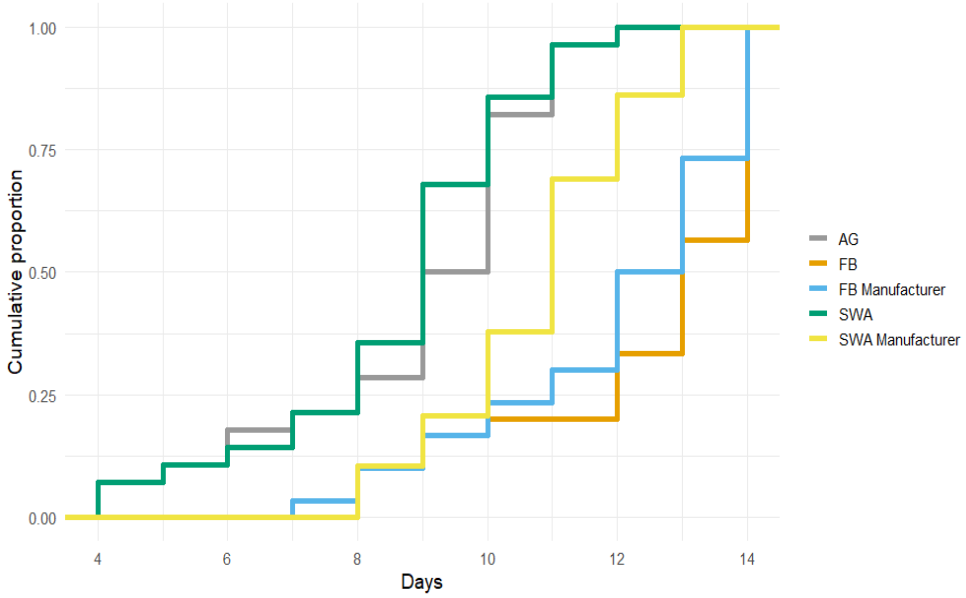



**Figure 8.3** An empirical cumulative distribution plot demonstrating the data availability for each of the reported models.

Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA).

### 8.3.3 Energy expenditure

Agreement statistics for TDEE are presented in table 8.3. In general, the FB manufacturer estimates were lower than the SWA Manufacturer estimates and algorithm predictions were higher. The FB manufacturer's estimation of EE had the best agreement with SWA in terms of RMSE, MAPE and mean difference. Figure 8.4 shows Bland-Altman plots for each model relative to the SWA. Most of the developed models showed visual evidence of underestimates at the lower end of the means of measures and overestimates at higher means of measures. The distribution of TDEE for each predictive model is shown in histograms in figure 8.5, which demonstrates comparable distributions for most estimation methods, with the means shifted upwards for all the machine learning models. Notably, the AG model produced estimates of TDEE higher than any other model.

**Table 8.3** Equivalence and agreement statistics for algorithms relative to the SenseWear armband for TDEE.

Equivalence refers to statistical equivalence tests ($p<0.05$). Abbreviations: Root mean squared error (RMSE), Mean absolute percentage error (MAPE). ActiGraph (AG), Fitbit (FB), SenseWear (SWA).

| | n | Predicted (kcal/day) | SWA Manufacturer (kcal/day) | MAPE | RMSE | Equivalence |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | 28 | 3291.14 ± 787.21 | 2906.56 ± 487.52 | 16.65 | 635.52 | |
| **AG Neural Network** | 28 | 3303.4 ± 811.52 | 2906.56 ± 487.52 | 16.87 | 663.59 | |
| **AG Random Forest** | 28 | 3288.51 ± 774.87 | 2906.56 ± 487.52 | 17.05 | 637.48 | |
| **FB Gradient Boost** | 29 | 3067.92 ± 716.9 | 2927.94 ± 492.38 | 13.03 | 494.97 | |
| **FB Manufacturer** | 29 | 2798.91 ± 611.71 | 2927.94 ± 492.38 | 8.98 | 287.71 | Equivalent |
| **FB Neural Network** | 29 | 3028.64 ± 723.3 | 2927.94 ± 492.38 | 12.94 | 480.93 | Equivalent |
| **FB Random Forest** | 29 | 3091.03 ± 731.21 | 2927.94 ± 492.38 | 13.52 | 519.00 | |
| **SWA Gradient boost** | 28 | 3089.7 ± 698.41 | 2906.56 ± 487.52 | 13.47 | 486.49 | |
| **SWA Neural Network** | 28 | 3074.88 ± 694.06 | 2906.56 ± 487.52 | 13.62 | 490.43 | |
| **SWA Random forest** | 28 | 3104.76 ± 697.89 | 2906.56 ± 487.52 | 14.17 | 504.07 | |

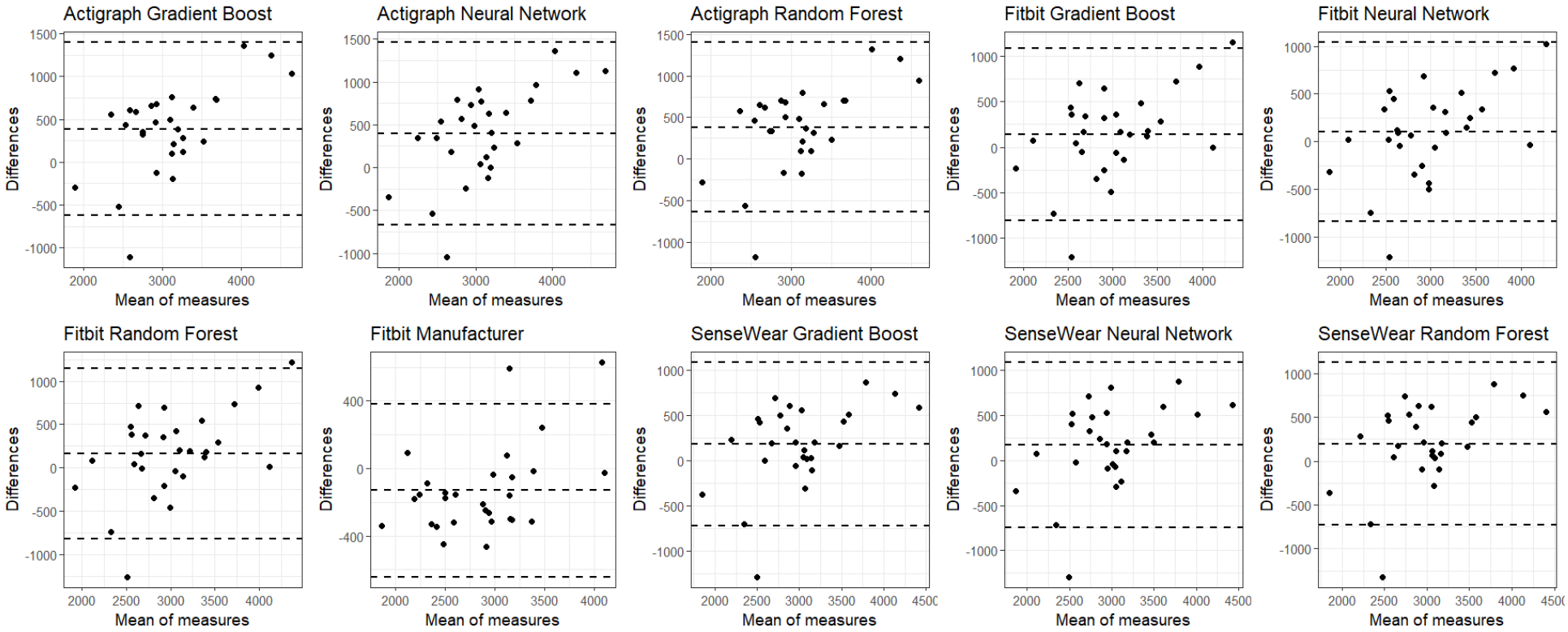



**Figure 8.4** Bland-Altman plots detailing the differences between the respective models and the SenseWear armband for total daily energy expenditure (kcal/day).

Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA)


**Figure 8.5** Bland-Histograms detailing the distribution of TDEE (kcal/day) for each of the models.

### 8.3.3.1 BMI & TDEE analysis

The boxplots in Figure 8.6 shows each model's predictions per tertile of BMI and TDEE. For all comparisons of BMI and TDEE, the AG models produced a higher estimate of TDEE than the SWA manufacturer.

The deviations (SWA TDEE – Model TDEE) closest to 0 for the FB models occurred in the low BMI group (all models deviating by less than 10 kcal) and for the SWA models it was the SWA random forest (overestimation of 23 kcal/day) in the low BMI group.

For FB models, the largest mean difference in TDEE was seen in the low TDEE groups where the FB models overestimated by between 269 (Random Forest) and 189 (Neural network) kcal/day. In the high TDEE group, where overestimation was between 190 (Neural network) and 250 (Random Forest) kcal/day and in the high BMI group, where overestimates were between 159 (Neural network) and 264 (Random Forest). Similar patterns and directions in these deviations were seen in the SWA estimates, where the largest overestimate was the SWA random forest, in the high TDEE comparison (364 kcal/day).

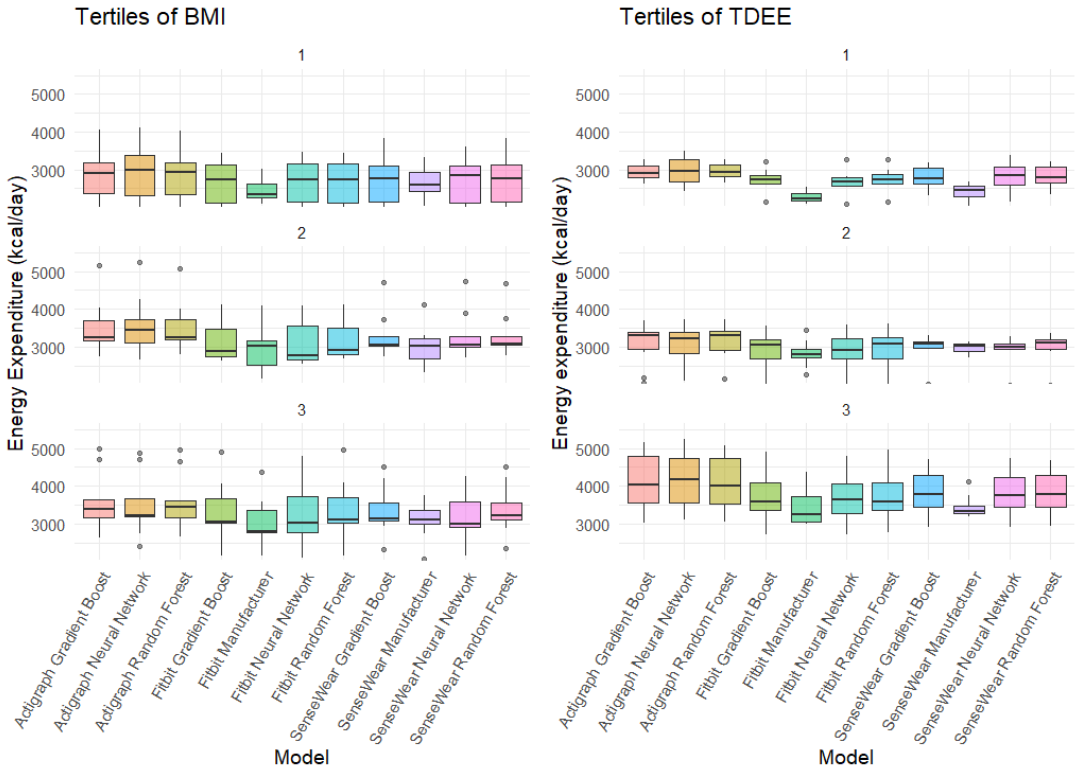



**Figure 8.6** Boxplots detailing the total daily energy expenditure for each of the models split by BMI (left) and TDEE (right) tertiles.

The top row (1) refers to the lowest BMI and TDEE and bottom (3) refers to the largest. The SWA manufacturer estimates can be seen in the ‘SenseWear’ group. Data are presented for participants for whom TDEE could be approximated by all models.

### 8.3.3.2 Sensitivity analysis: DIT

The effect of scaling DIT by different values is visually represented in figure 8.7. Above a ~5% DIT it appears that the AG/SWA models produce higher estimates of TDEE that the SWA manufacturer and at around 7.5% DIT adjustment all models were higher than the mean SWA manufacturer estimate.

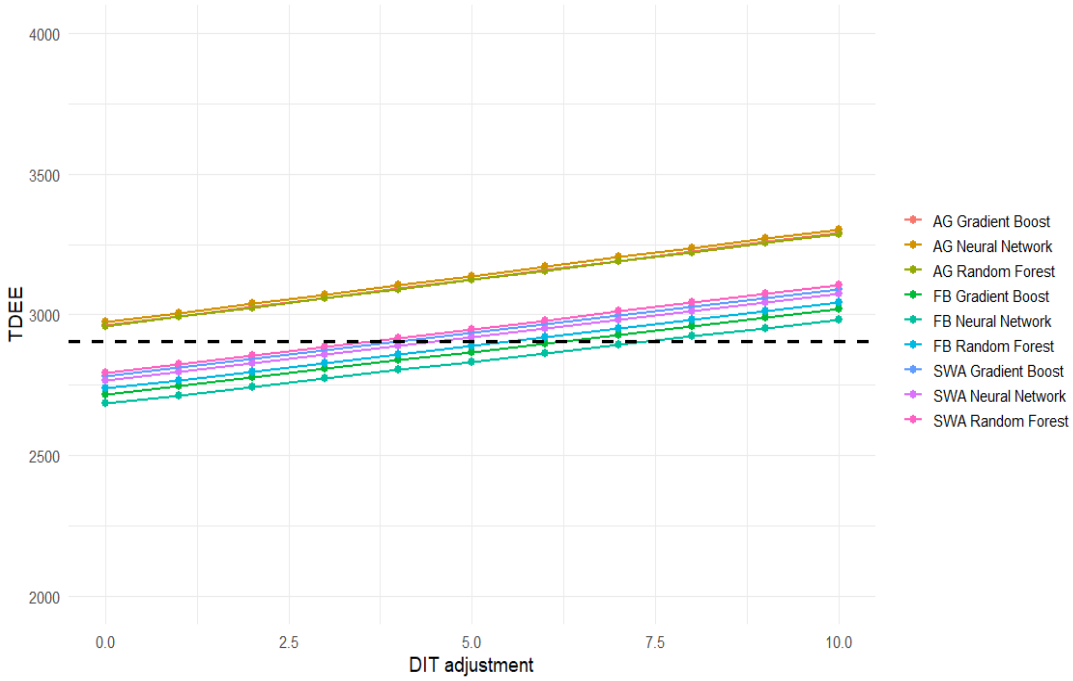



**Figure 8.7** A figure representing the effect of different DIT estimates on the final TDEE outcomes.

The black dashed line represents the mean estimate of the SWA armband and TDEE is presented in kcal/day. Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA). Gradient boost (GB), Neural network (NN), Random Forest (RF), Total daily energy expenditure (TDEE), Dietary induced thermogenesis (DIT).

### 8.3.3.3 Sensitivity analysis: Outlier removal

It was noted that some participants removed the device for comfort reasons at night-time. These participants appeared to have higher PAL values when compared to the model predictions. As the SWA manufacturer algorithm is proprietary it is not possible to determine how the algorithm behaves in this instance. As such, table 8.4 reports the TDEE agreement statistics with

those participants (n=2) removed. When compared to table 8.3 it is evident that the MAPE values tend to decrease for most estimators, but the FB manufacturer remains in closest agreement to the SWA.

**Table 8.4** Equivalence and agreement statistics for algorithms relative to the SenseWear armband for TDEE after removing potential outliers.

Equivalence refers to statistical equivalence tests ($p<0.05$). Abbreviations: Root mean squared error (RMSE), Mean absolute percentage error (MAPE). ActiGraph (AG), Fitbit (FB), SenseWear (SWA).

| | n | Predicted (kcal/day) | SWA Manufacturer (kcal/day) | MAPE | RMSE | Equivalence |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | 26 | 3382.1 ± 740.53 | 2905.48 ± 502.79 | 14.93 | 614.56 | |
| **AG Neural Network** | 26 | 3393.54 ± 769.99 | 2905.48 ± 502.79 | 15.31 | 649.22 | |
| **AG Random Forest** | 26 | 3383.41 ± 718.83 | 2905.48 ± 502.79 | 15.06 | 610.45 | |
| **FB Gradient Boost** | 27 | 3150.38 ± 671.62 | 2928.48 ± 507.3 | 10.32 | 435.79 | |
| **FB Manufacturer** | 27 | 2823.59 ± 624.74 | 2928.48 ± 507.3 | 8.27 | 271.25 | Equivalent |
| **FB Neural Network** | 27 | 3109 ± 682.76 | 2928.48 ± 507.3 | 10.18 | 416.98 | |
| **FB Random Forest** | 27 | 3177.74 ± 680.05 | 2928.48 ± 507.3 | 10.64 | 458.49 | |
| **SWA Gradient boost** | 26 | 3178.93 ± 640.69 | 2905.48 ± 502.79 | 10.51 | 415.75 | |
| **SWA Neural Network** | 26 | 3164.22 ± 635.37 | 2905.48 ± 502.79 | 10.57 | 417.97 | |
| **SWA Random forest** | 26 | 3197.41 ± 632.88 | 2905.48 ± 502.79 | 11.07 | 432.10 | |

#### 8.3.3.4 Sensitivity analysis: Predicted RMR

A further sensitivity analysis explored the effect of utilising the WHO (Schofield) RMR equations (Miller et al., 2013; Schofield, 1985; World Health Organization, 1985) in each of the predictive models rather than the measured RMR. This was conducted to investigate the extent to which the WHO RMR value used in the SWA contributes to the observed differences. The FB manufacturer estimates are not included in these results because it is not clear which RMR equation is used. One participant is also excluded here as RMR data were not available. A paired t-test revealed a significant difference between the two estimates ($t = 3.27$, $df = 28$, p-value = 0.003), with the RMR values used in this study being 161 kcal/day higher on average.

Models utilising the predicted rather than measured RMR values are shown in table 8.5. Only the AG Neural Network was not equivalent, and the mean TDEE for all other models decreased. The MAPE and RMSE fall notably when compared to the comparisons presented in table 8.3 with percentage decreases in RMSE reaching 63% (AG random forest). Figure 8.8 demonstrates the agreement between the WHO predicted RMR and the values used in this study. Visually, evidence of an underestimate of RMR by the WHO equation at the higher end of RMR, and an overestimate of RMR by the WHO equation at the lower end of RMR was observed.

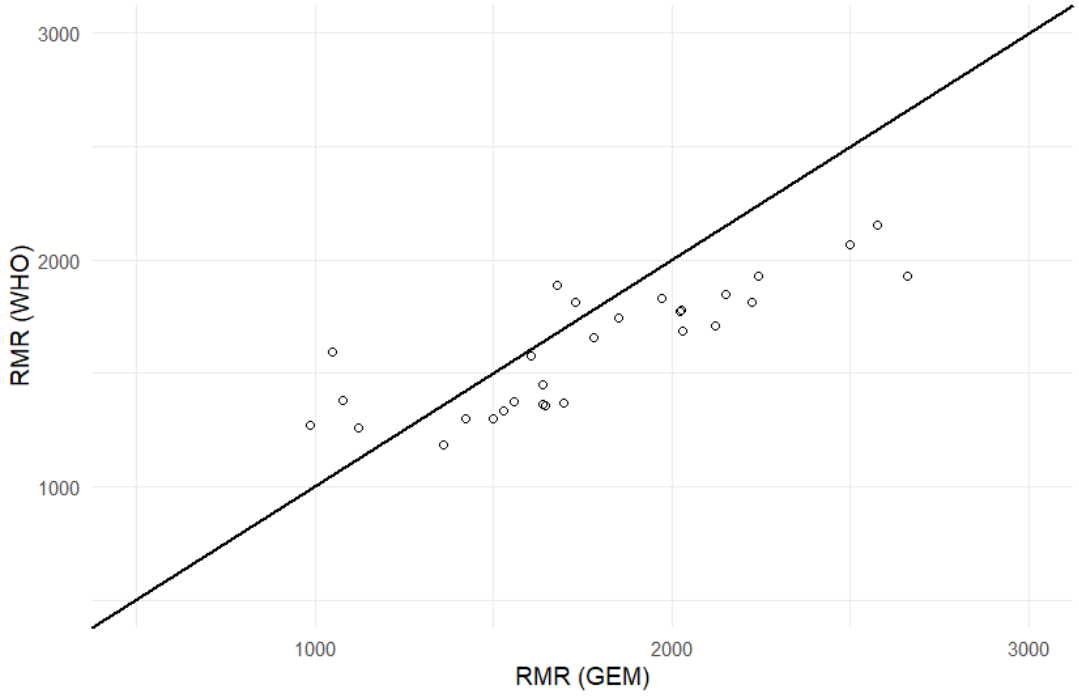



**Figure 8.8** A comparison between the RMR values used in this study and those predicted by the WHO equation.

RMR (GEM) refers to the measured RMR values and RMR (WHO) refers to values predicted by the RMR equation. A line of identity represents y=x.

**Table 8.5** Equivalence and agreement statistics for algorithms relative to the SenseWear for TDEE utilising predicted RMR.

RMR is predicted by the WHO equations. Equivalence refers to statistical equivalence tests ($p<0.05$). Abbreviations: Root mean squared error (RMSE), Mean absolute percentage error (MAPE).

| | n | Predicted (kcal/day) | SWA Manufacturer (kcal/day) | MAPE | RMSE | Equivalence |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | 28 | 3015.07 ± 540.51 | 2906.56 ± 487.52 | 6.27 | 238.63 | Equivalent |
| **AG Neural Network** | 28 | 3025.43 ± 564.09 | 2906.56 ± 487.52 | 7.25 | 270.90 | Equivalent |
| **AG Random Forest** | 28 | 3010.45 ± 519.6 | 2906.56 ± 487.52 | 6.35 | 235.10 | Equivalent |
| **FB Gradient Boost** | 29 | 2783.33 ± 452.86 | 2927.94 ± 492.38 | 7.54 | 267.48 | Equivalent |
| **FB Neural Network** | 29 | 2747.07 ± 462.83 | 2927.94 ± 492.38 | 8.64 | 282.18 | Equivalent |
| **FB Random Forest** | 29 | 2801.96 ± 457.74 | 2927.94 ± 492.38 | 7.45 | 263.02 | Equivalent |
| **SWA Gradient boost** | 28 | 2831.03 ± 477.88 | 2906.56 ± 487.52 | 6.69 | 221.10 | Equivalent |
| **SWA Neural Network** | 28 | 2816.68 ± 470.88 | 2906.56 ± 487.52 | 7.39 | 240.23 | Equivalent |
| **SWA Random forest** | 28 | 2842.83 ± 467.87 | 2906.56 ± 487.52 | 6.87 | 222.09 | Equivalent |

### 8.3.3.5 Patterns in EE estimates

Patterns amongst the different models of EE estimation are shown in the pairs plot in figure 8.9. Comparisons are made between models for sedentary, light and MVPA represented by green, yellow and red, respectively. Between model comparisons overall showed high correlations ($r > 0.88$). The greatest associations for the algorithms tended to be within a device (e.g. FB random forest, FB gradient boost and FB neural network).

The SWA manufacturer was most highly correlated with the FB manufacturer estimates (r=.915 to r=.947) for specific activities.

Next, figure 8.10 shows a density plot of subject-level Pearson's correlations between the EE predictions in kcal/min and the heart rate measure by the Polar heart rate monitor. Notably, the distribution of correlations between the SWA manufacturer EE predictions and heart rate trends closer to 0. The SWA and AG (Random forest and gradient boost) models appear to have the greatest correlations, with the peak of the distribution closest to 1, however, this is to be expected because the Polar heart rate is a predictive feature in the SWA and AG models. Lastly for EE, the behaviour of each of the models is shown by a time-series plot in figure 8.11. This represents the minute level predictions of each estimator for a single day and participant. A notable pattern in the AG models is the frequency of spikes in EE at 5-7 kcal/min, where other models are estimating the subject is closer to resting EE.

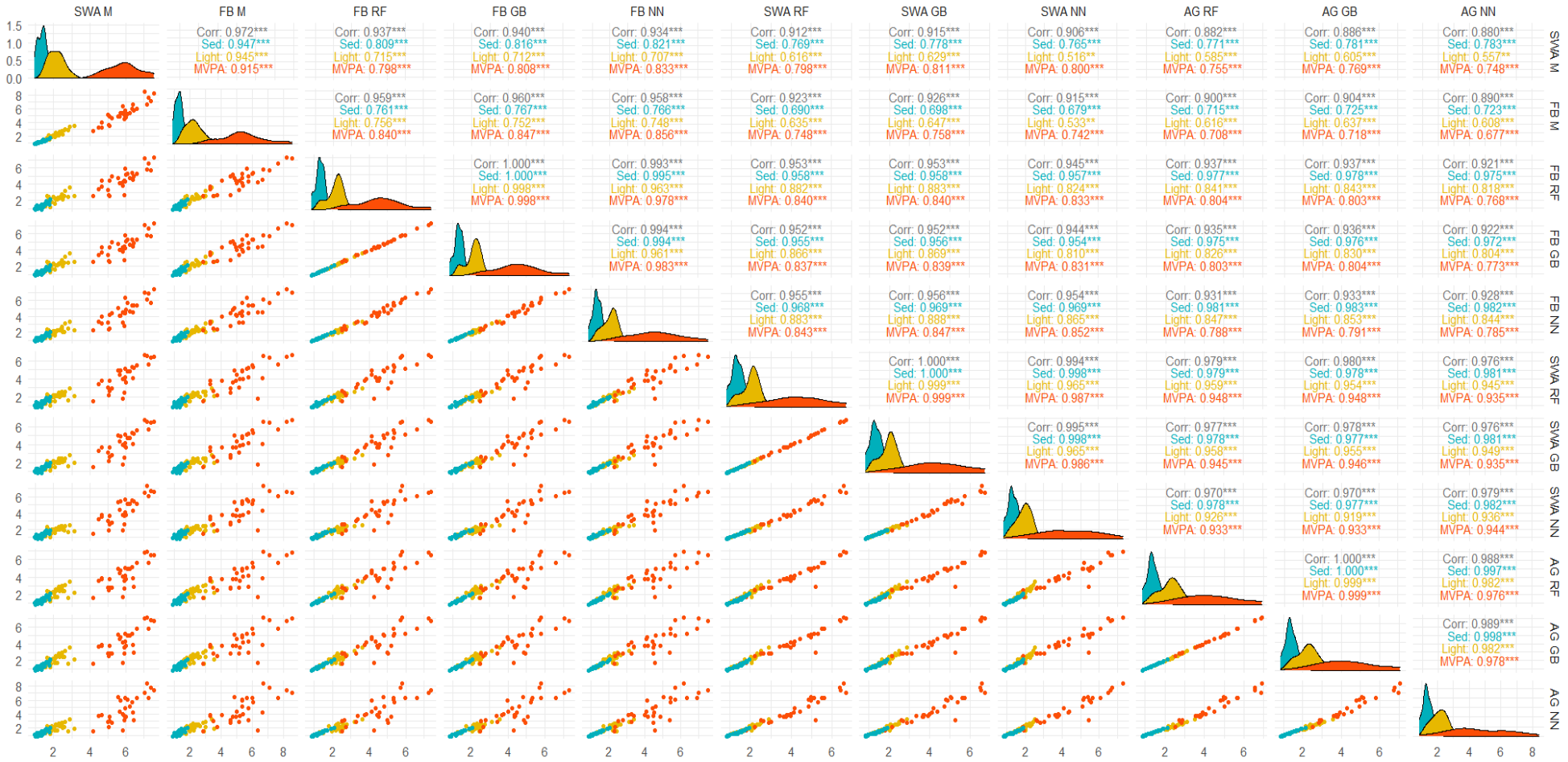



**Figure 8.9** A pairs plot demonstrating the associations between the models tested in this study.

Data are shown in kcal/min and coloured by activity with sedentary = green, light = yellow and MVPA= red. The bottom left plots are scatter plots for all paired comparisons, the diagonal plots are density plots, displaying the distribution of each model's predictions and the upper right panels represent pairwise Pearson's correlations (r). Abbreviations: Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA). Gradient boost (GB), Neural network (NN), Manufacturer (M), Random Forest (RF). Sedentary (Sed), moderate to vigorous physical activity (MVPA).

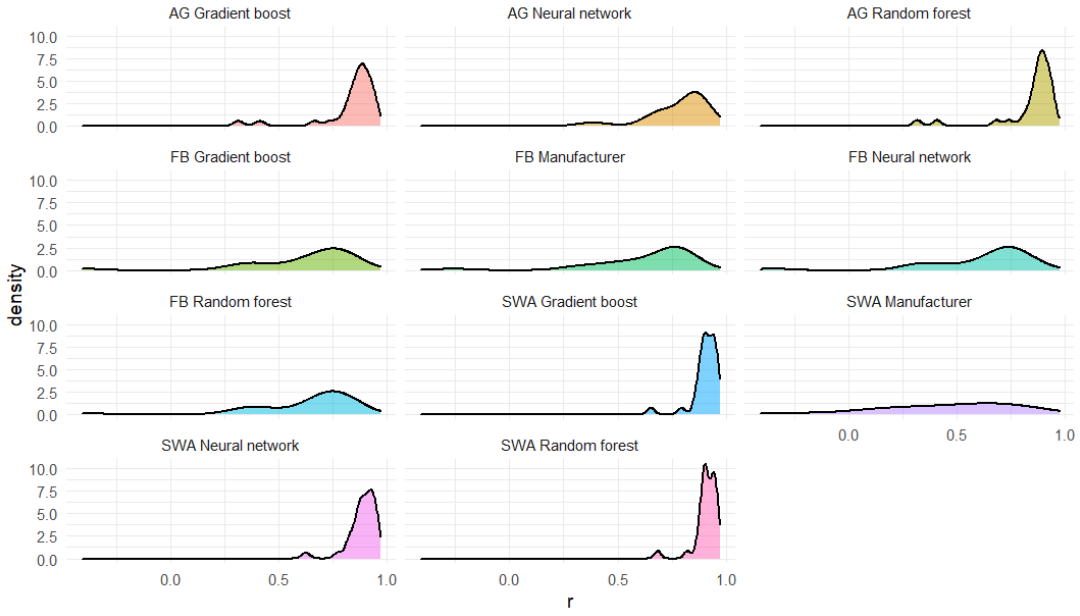



**Figure 8.10** Density plots demonstrating the distribution of the individual level correlations between heart rate (Polar) and EE predictions.

The plot is faceted by each of the included models and correlations are calculated on minutes classified as 'MVPA' by the SWA. Abbreviations: Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA), Moderate to vigorous physical activity (MVPA).


**Figure 8.11** A time series plot of minute level EE for a random subject and day (4 am – 10 pm) for each of the included models.

### 8.3.4 Physical activity level

Agreement statistics for PAL are presented in table 8.6, with RMSE and MAPE for each of the predictions relative to the SWA manufacturer. The FB manufacturer estimates were lower than estimates of the SWA manufacturer but equivalence between these measures was observed. Figure 8.12 shows Bland-Altman plots for each model relative to the SWA manufacturer. It is important to note a large deviation for two points at > 2 PAL, and aside from these outliers, most points for most models fall within limits of agreement. The distribution of PAL for each model is shown in histograms in figure 8.13. A notable difference is that the SWA manufacturer and FB manufacturer have a lower mean PAL than the machine learning models and that the SWA manufacturer has two participants between 2.5 and 3, which is not seen for the other models.

**Table 8.6** Equivalence and agreement statistics for algorithms relative to the SenseWear armband for PAL.

Equivalence refers to statistical equivalence tests ($p<0.05$). Abbreviations: Root mean squared error (RMSE), Mean absolute percentage error (MAPE). Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA).

| | n | Predicted | SWA Manufacturer | MAPE | RMSE | Equivalence |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | 28 | 1.87 ± 0.2 | 1.7 ± 0.38 | 16.65 | 0.38 | |
| **AG Neural Network** | 28 | 1.88 ± 0.24 | 1.7 ± 0.38 | 16.87 | 0.39 | |
| **AG Random Forest** | 28 | 1.87 ± 0.19 | 1.7 ± 0.38 | 17.05 | 0.39 | |
| **FB Gradient Boost** | 29 | 1.72 ± 0.17 | 1.69 ± 0.38 | 13.03 | 0.33 | Equivalent |
| **FB Manufacturer** | 29 | 1.59 ± 0.29 | 1.69 ± 0.38 | 8.98 | 0.19 | Equivalent |
| **FB Neural Network** | 29 | 1.7 ± 0.18 | 1.69 ± 0.38 | 12.94 | 0.33 | Equivalent |
| **FB Random Forest** | 29 | 1.73 ± 0.17 | 1.69 ± 0.38 | 13.52 | 0.34 | Equivalent |
| **SWA Gradient boost** | 28 | 1.76 ± 0.19 | 1.7 ± 0.38 | 13.47 | 0.35 | |
| **SWA Neural Network** | 28 | 1.76 ± 0.21 | 1.7 ± 0.38 | 13.62 | 0.35 | Equivalent |
| **SWA Random forest** | 28 | 1.77 ± 0.19 | 1.7 ± 0.38 | 14.17 | 0.36 | |

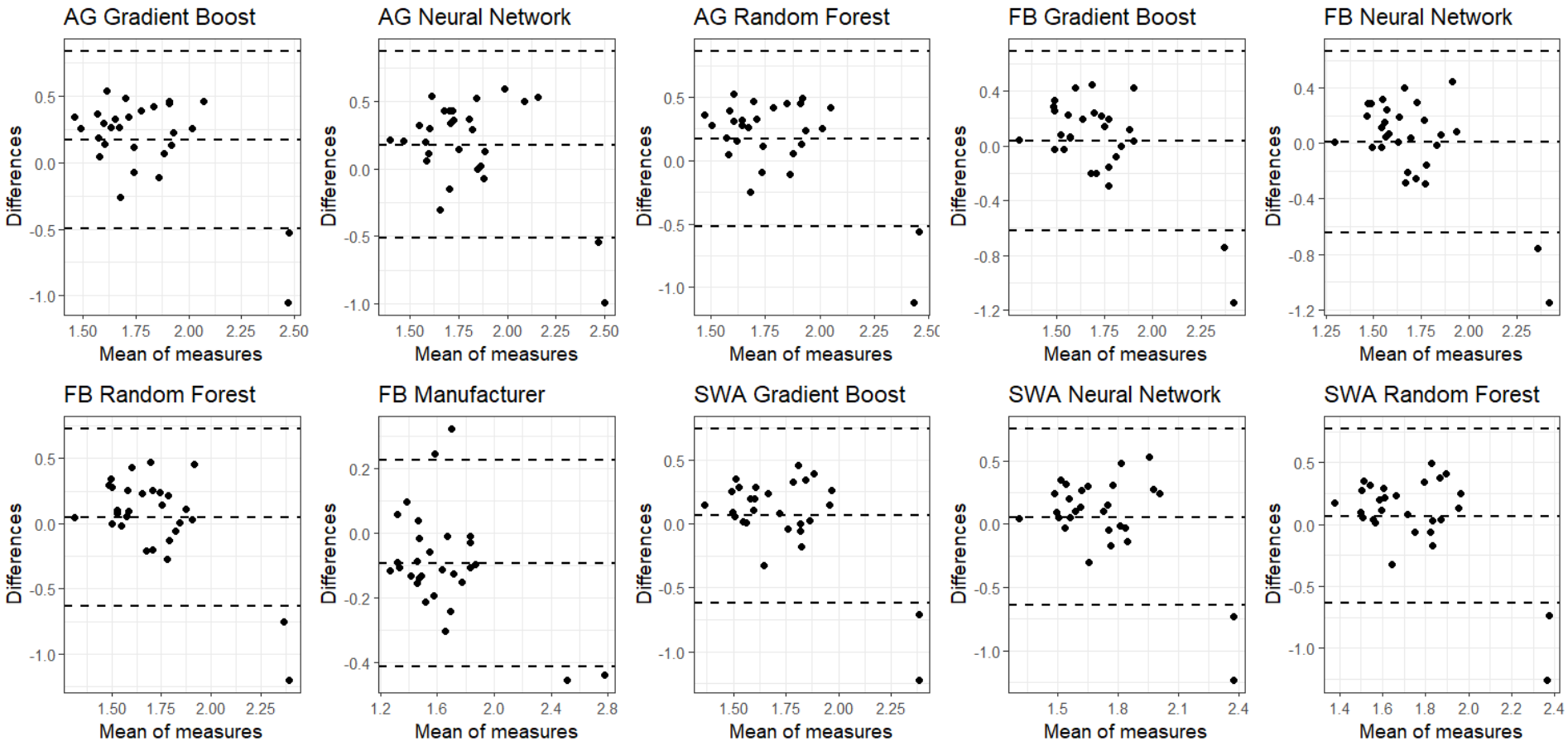



**Figure 8.12** Bland-Altman plots detailing the differences between the respective models and the SWA for PAL.

Abbreviations: ActiGraph (AG), Fitbit (FB), SenseWear (SWA)

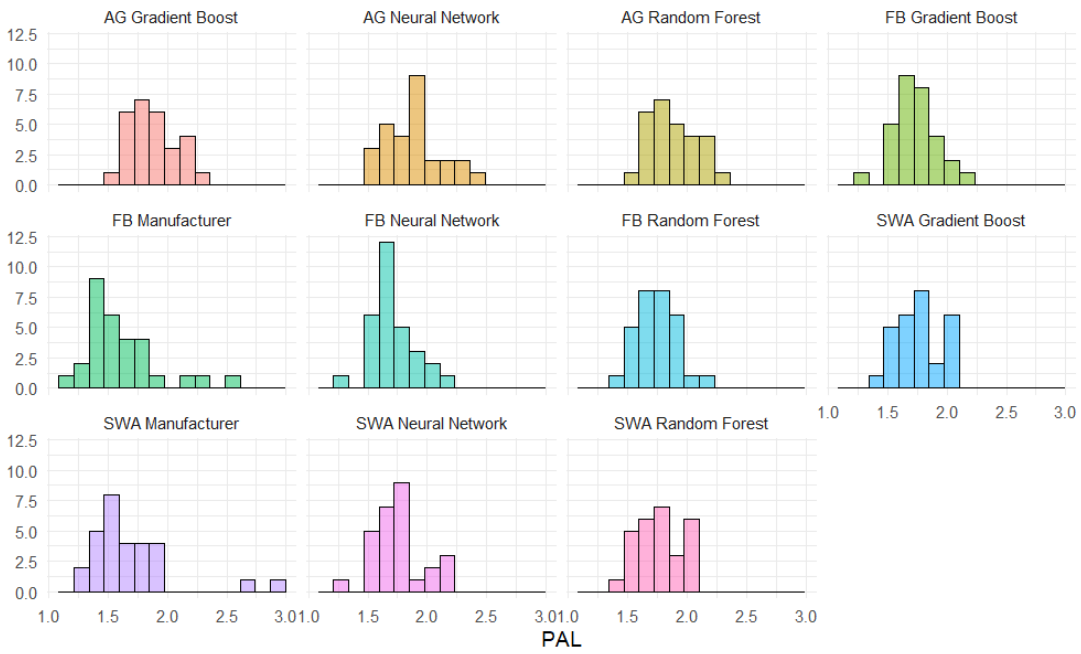



**Figure 8.13** Histograms detailing the distribution of PAL for each of the models.

### 8.3.5 Energy intake

Agreement statistics for EI are presented in table 8.7. All the developed machine learning models resulted in a mean EI value which was higher than the SWA manufacturer, whereas the Fitbit manufacturer estimates were lower. In appendix 4.2.1, Bland-Altman plots for each model relative to the SWA manufacturer are presented. Visually, it was apparent that the biggest differences for most models were at the lower and upper mean of measures, which indicates proportional bias. The distribution of EI for each predictive model is shown in histograms in appendix 4.2.2. The AG models have higher mean than the other models, as attributable to the high TDEE estimates of these models.

**Table 8.7** Equivalence and agreement statistics for algorithms relative to the SenseWear armband for energy intake (kcal/day).

Equivalence refers to statistical equivalence tests (p<0.05).
Abbreviations: Root mean squared error (RMSE), Mean absolute percentage error (MAPE).

| | n | Predicted (kcal/day) | SWA Manufacturer (kcal/day) | MAPE | RMSE | Equivalence |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | 28 | 3141.7 ± 1085.41 | 2757.13 ± 871.71 | 18.07 | 635.52 | |
| **AG Neural Network** | 28 | 3153.97 ± 1103.81 | 2757.13 ± 871.71 | 17.99 | 663.59 | |
| **AG Random Forest** | 28 | 3139.08 ± 1071.6 | 2757.13 ± 871.71 | 18.53 | 637.48 | |
| **FB Gradient Boost** | 29 | 2882.05 ± 996.64 | 2742.06 ± 859.84 | 14.14 | 494.97 | |
| **FB Manufacturer** | 29 | 2613.03 ± 948.03 | 2742.06 ± 859.84 | 9.52 | 287.71 | Equivalent |
| **FB Neural Network** | 29 | 2842.77 ± 997.41 | 2742.06 ± 859.84 | 14.40 | 480.93 | Equivalent |
| **FB Random Forest** | 29 | 2905.15 ± 1004.51 | 2742.06 ± 859.84 | 14.71 | 519.00 | |
| **SWA Gradient Boost** | 28 | 2940.27 ± 1024.64 | 2757.13 ± 871.71 | 14.44 | 486.49 | |
| **SWA Neural Network** | 28 | 2925.45 ± 1017.97 | 2757.13 ± 871.71 | 14.75 | 490.43 | |
| **SWA Random Forest** | 28 | 2955.33 ± 1019.62 | 2757.13 ± 871.71 | 15.24 | 504.07 | |

## 8.4 Discussion

This chapter aimed to evaluate TDEE predictions from a series of machine learning models in a free-living environment. The results show that TDEE estimates were, on average, higher than the SWA manufacturer algorithm and deviated from these estimates by more than FB manufacturer estimates, as measured by RMSE and MAPE. Within devices, machine learning models produced similar results, and this was particularly true of the tree-based algorithms (random forest and gradient boost). When these TDEE estimates were used to derive EI and PAL, similar patterns were observed, and this is expected given the centrality of the TDEE estimates to these calculations. Sensitivity analyses highlighted the potentially large differences in TDEE estimates attributable to the use of the WHO RMR equations and how the disparities between the developed algorithms and the SWA is often greatest in the highest BMI and TDEE groups.

It is important to consider the potential reasons for the difference in estimates of TDEE (and therefore PAL and EI) between the SWA and the models used in this study. One reason may be the addition of 10% to the sum of model predictions to account for the energetic cost of digestive processes (Westerterp et al. 2004). It is unclear whether DIT adjustments are made in the SWA manufacturer model. If the calibration data was collected when the subjects were in the fed state then it would be unnecessary to make such an adjustment (Löf et al., 2013). This highlights an issue with the use of proprietary algorithms. If researchers are interested in estimating PAEE in a group with an atypical rate of EE attributable to DIT (e.g., high protein or alcohol intake (Westerterp et al. 2004)), considerable uncertainty would exist in TDEE, DIT and PAEE estimates. By contrast, when details regarding the calibration data and model assumptions are provided, adjustments could be more precisely adapted to specific experimental subjects.

When compared within devices the machine learning models produce similar estimates of EE, for example, the FB random forest performs similarly to the FB gradient boost. The AG models tended to overestimate more notably. In **chapter 7**, the potential for models to overfit the training data was discussed, which will ultimately lead to poorer performance when a model is used to predict an outcome on a new dataset, and this could be a source of error. The data being used for predictions in this study was collected in a free-living environment, where the distribution of movement is likely to differ from the training data and this disparity could negatively influence prediction

accuracy (Kuhn and Johnson 2013, pp 61-64). The large number of accelerometer variables in the AG models (detailed in **chapter 7**) creates a more complex model, which also could lead to overfitting, and this is an issue of neural networks in particular (Hastie, Tibshirani, and Friedman 2009. pp 398). Variable selection methods specific to neural networks (May et al., 2011) and methodologies such as dropout (Srivastava et al., 2014) may assist in addressing this in future research in this field. In any case, there is no guarantee that a particular algorithm will perform best (Wolpert & Macready, 1997) and this highlights the advantage of testing numerous algorithms within devices.

It may be possible that the SWA underestimates RMR (and therefore TDEE). The RMR equation used by the SWA software is the WHO equation and these models have been criticised as being inaccurate previously, for example in males aged 40-49 and 50-59 the WHO equation gives an error in RMR of more than 6% (Müller et al., 2004). This is not surprising given that the model is a linear function of bodyweight, with separate models depending on gender and age (Rao et al., 2012). It is thought that RMR may be more appropriately described as a power law relative to mass, i.e. $\mathrm{RMR} = \mathrm{a} \times \mathrm{Weight}^{\mathrm{b}}$ where a is a coefficient and b is a power coefficient (Heymsfield et al., 2012; Kleiber, 1947; Livingston & Kohlstadt, 2005). This means that linear equations are unlikely to predict RMR accurately across a wide range of subjects. Indeed, the scatter plot comparing the two RMR estimates show deviations from the y=x line at the lower and upper ends of RMR. The mean RMR estimated by the WHO equation was 1611 ± 271kcal/day compared to the measured value of 1772 ± 443 kcal/day, and 11 subjects differed by at least ± 300 kcal/day, with a maximum difference of 730 kcal/day. This observation is important in the context of the relatively wide limits of agreement observed in the Bland-Altman analyses. A sensitivity analysis directly investigated this hypothesis and a far closer agreement was achieved when predicted rather than measured RMR was used to convert METs to kcals, indicating that the WHO equations are indeed a critically important factor in the TDEE estimates of the SWA. This may not present as an issue in some studies, where the sample is reflective of the development cohort for the WHO equations but it is clearly imperative in the present sample. An alternative explanation here may relate to inaccuracy in the GEM indirect calorimeter, which is taken to be the gold-standard measure. Indeed, compared to a reference Deltatrac II Metabolic Monitor (Datex-Ohmeda Inc.), one study reported that the GEM significantly overestimated EE, despite a high degree of repeatability in healthy subjects (

Kennedy et al., 2014). Nonetheless, the analyses presented here provide strong evidence that RMR estimates play a large role in determining model differences.

The fact that the greatest difference between the SWA and the machine learning algorithms was seen in the highest TDEE group is enlightening in the context of the previous literature. A recent review reported that the manufacturer estimates of the SWA produce valid estimates of EE in the general population, but consistent underestimations in more athletic populations, with higher rates of EE (Koehler & Drenowatz, 2017). This chapter reports steps/day for the entire sample and each of the TDEE/BMI groups to use as an indicator of physical activity (ambulatory only) and to facilitate comparisons with other studies utilising the same device. The study of Shook, *et al.* also took a tertiled approach to their analysis. Their highest TDEE group had an RMR of 1690 ± 277 kcal/day, averaged 8138 ± 3011 steps/day and their TDEE value (measured by DLW) was 3170 ± 519 kcal/day and the SWA underestimated TDEE by ~160 kcal (Shook et al., 2018). In the present study, the sample RMR was 79 kcal/day higher and averaged nearly 3900 steps/day more. Most of the predictive models overestimated TDEE relative to the SWA. Based on this observation, it would be reasonable to assume that the DLW estimates would be comparable or higher than the 3170 kcal/day value reported by Shook and colleagues.

Despite the lack of DLW data (discussed in limitations below), this work has explored intra-day relationships between models, devices, and algorithms, which DLW will not be able to provide. In the time series plot of intra-day predictions, the patterns of EE are extremely similar between models, with some being slightly more sensitive and producing more frequent ‘spikes’ in EE estimates. Correlational analyses indicated that the SWA and FB manufacturer estimates are more closely related than any of the developed machine learning algorithms, which are more closely associated with each other. Given that the machine learning algorithms were trained on very similar datasets, this is easy to understand. It is likely that the lack of a heart rate measure in the SWA is also a central factor in creating differences between the developed algorithm estimates and the SWA. It is known that heart rate and EE are tightly related during physical activity (Bonomi et al. 2015; Brage et al. 2007; O’Driscoll, Turicchi, Beaulieu, et al. 2020; O’Driscoll, Turicchi, Hopkins, et al. 2020; Ceesay et al. 1989; Ekelund et al. 2002) and in **chapter 7** this was confirmed by showing that the machine

learning models put significant emphasis on heart rate variables for the prediction of EE. By correlating each model's predicted EE with measured heart rate during MVPA (as determined by the SWA), it was shown that the SWA manufacturer's EE predictions are most weakly related to measured heart rate. **Chapter 4, 5** and **7** show that whilst the SWA is accurate on aggregate, it does not accurately predict the EE of all activities, and this is particularly true in non-ambulatory activities. By contrast, the machine learning algorithms appear to be capable of modelling a wide range of activities, albeit in a controlled laboratory. Again, for these reasons, the SWA must not be considered to be a criterion in this analysis.

Some advantages and limitations of this work must be discussed. A significant advantage is the free-living nature of this study, which is the first to explore the validity of machine learning models in a truly free-living setting (up to 14 days, depending on data availability). A clear limitation of this work is the temporary lack of DLW data due to COVID-19, which forced comparison to the TDEE predictions of the SWA. The SWA is arguably one of the most accurate wearables for the estimation of TDEE in the general population (O'Driscoll, Turicchi, Hopkins, et al. 2020; Shook et al. 2018; O'Driscoll, Turicchi, Beaulieu, et al. 2020) but it is not a criterion measure. For example, previous work has considered the individually calibrated Actiheart device to be a criterion measure rather than the BodyMedia Core (a similar device from the same manufacturer) (Chowdhury et al., 2017). A further limitation of this study lies in the calibration data. Whilst the training data did incorporate a relatively wide and representative set of activities, simulating truly naturalistic behaviours in a laboratory is a challenge. It was therefore necessary to make assumptions about the EE of subjects outside the range of the training data (i.e. prolonged sedentariness, sleeping), which was identified by a classification algorithm. Previous studies have shown a tendency for models to overpredict sedentary activities (Montoye et al. 2015; Staudenmayer et al. 2015; O'Driscoll, Turicchi, Hopkins, et al. 2020) and as western adults spend the majority of their day in sleeping/sedentary behaviours, this has the potential to result in large overestimates in TDEE. It is anticipated that protocols involving whole room calorimeters will allow further improvement of predictive capabilities, particularly in extreme sedentariness and sleeping behaviours, rather than the sampling approach taken here.

Another potential limitation is the assumption of 10% DIT for all subjects. In the methods, attempts were made to gain a quantitative understanding of

the ‘worst-case scenario’ related to the protocols used to collect calibration data, based on a model reported in a previous study (Reed & Hill, 1996). Furthermore, estimates of each model on a continuum from 0% to 10% DIT are reported, where most models trended towards and then below the average SWA estimate as DIT estimates go towards 0%. In reality, DIT estimates of 10% are a very rough approximation and vary widely based on nutrient and alcohol intake (Westerterp et al. 2004) and perhaps body composition and activity status (de Jonge & Bray, 1997). Large variation may also exist within-subjects, with variation in DIT from ventilated hood assessments and whole room calorimeter studies suggesting a wide variation (Ravussin et al., 1986; Segal et al., 1992; Tataranni et al., 1995; Weststrate, 1993). Future work should investigate the predictors of DIT under controlled conditions (i.e., whole room calorimeters) and consider how this could refine TDEE prediction models.

Next, the composition of weight change was approximated with the BodPod. Whilst this is a widely used tool for two-compartment body composition measurements, with a high degree of precision and accuracy (Fields et al., 2002), it may not compare to the accuracy of other measurement tools such as DEXA (Racette et al., 2012). Therefore, the possibility remains that this measure adds bias to the EI assessments, although there is no reason to expect that this would disproportionately affect any one of the models. Also, a two-compartment model of body composition considers FFM to be a single, uniform compartment and cannot account for the non-energetic fluctuations in weight (i.e. hydration, total body water changes), which could fluctuate markedly before or between the two measurement points in this study (Bhutani et al., 2017). Lastly, caution must be exercised in extending these results to different populations. The sample in this study was highly active (as measured by steps/day) with steps nearly double previously reported values in some European (Althoff et al., 2017), and North American adult populations (Bassett et al., 2010; Tudor-Locke et al., 2009). Furthermore, the percentage of FM (21.7%) is considerably lower than might be expected in the general population, for example, the Biobank study reports 24.4% and 35.5% FM for men and women, respectively (Bradbury et al., 2017). It is of great importance to evaluate and perhaps refine these models for use in different populations.

## 8.5 Conclusion

This study has presented estimates of TDEE from novel hierarchical machine learning models in a group of healthy adults. In general, the presented models overestimated TDEE, PAL and EI relative to the SWA and this was particularly true in the AG models. Considering that the SWA has been demonstrated to underestimate in those with the highest rates of EE and that the RMR equations used by the SWA are established to have shortcomings, these results may be considered to be encouraging. Whilst the DLW data is pending caution must be exercised in stating that either the machine learning models, or the manufacturer estimates offer an advantage of accuracy. However, machine learning methods have a significant advantage in that the assumptions and techniques used to estimate EE are far more transparent.

# Chapter 9 – Modelling the components of energy balance in the NoHoW cohort.

## 9.1 Introduction

Physiological and behavioural responses to a reduction in body weight contribute to the high probability of weight regain after weight loss (Kraschnewski et al., 2010). A prolonged negative energy balance (resulting in weight loss) alters circulating hormones (e.g. leptin, CCK, ghrelin) (Chearskul et al., 2008; Cummings et al., 2002; Geldszus et al., 1996) and increases appetite (Keim et al., 1998; Sumithran et al., 2011). With weight loss, the rate of EE also tends to decrease (Leibel & Hirsch, 1984) and this occurs through several pathways: First, weight loss reduces FFM and therefore RMR (Rosenbaum & Leibel, 2010). Reduced body mass also decreases the absolute EE for weight-bearing physical activity and non-exercise thermogenesis is also thought to decrease (MacLean et al., 2011). Taken together, these factors make successful weight loss maintenance extremely challenging and atypical (Melby et al., 2017). To prevent weight regain as a subject transitions from a period of weight loss to a period of habitual maintenance, consistent and diligent self-monitoring of energy balance behaviours will likely be required, perhaps extending over many years (Klem et al., 2000; Stubbs et al., 2019).

Achieving weight loss maintenance is extremely challenging and many models have been proposed which aim to explain how bodyweight is regulated in humans. Debate exists around the extent to which body weight is indeed regulated although evidence consistently shows that physiological and behavioural responses are asymmetric, with defence against weight loss being far stronger than weight gain (Stubbs & Turicchi, 2021). Here, three of the more prevalent models are briefly discussed. First, the ‘set-point’ model (Kennedy, 1953) suggests that a particular level of adiposity is ‘defended’, and appears to be supported by the frequency with which people regain weight after weight loss. The set-point, however, centralises adiposity and gives little consideration to other factors (i.e. FFM, environment, etc) (Speakman et al., 2011) and is contradicted by the increasing prevalence of obesity worldwide (Agha & Agha, 2017). An alternative model called the ‘settling-point’ model has been proposed which also relates to body composition but does not define a single equilibrium around which

bodyweight is regulated. The settling-point model, therefore, permits a role for societal and environmental factors in bodyweight regulation (Speakman et al., 2002). Importantly, the set and settling point theories do not account for the interactions between environmental and genetic factors and their impact on EI and therefore bodyweight (Müller, Geisler, Heymsfield, et al., 2018). The 'general model of intake regulation' (de Castro & Plunkett, 2002) hypothesises that 'uncompensated' and 'compensated' factors contribute to feeding behaviours and bodyweight in humans. A 'compensated' factor is typically physiological and can influence EI and be influenced by EI. By contrast, 'uncompensated' factors, which are primarily societal or environmental in nature, can influence EI but is not influenced by EI. Rather than proposing that body weight is regulated around some predefined and unalterable set point, this model suggests that the point at which bodyweight is defended is changeable. After a change in weight (i.e. the weight loss required to participate in the NoHoW trial), a drive towards the restoration of the previous weight is typical. However, the likelihood of this weight becoming a new defended weight depends on genetic factors interacting with the compensated and uncompensated factors (de Castro, 2010). Whilst it is generally accepted that there is some control of bodyweight in humans and that biological and behavioural factors and interactions between them are implicated (Melby et al., 2017), a widely accepted framework does not yet exist (Müller, Geisler, Heymsfield, et al., 2018). It is obvious that at the most fundamental level, weight regain results from EI exceeding EE. As such, the ability to measure EE and EI continuously would be an important development in elucidating the physiological and behavioural correlates of long-term weight outcomes after a period of weight loss (MacLean et al., 2015) and may allow for models of bodyweight regulation to be more rigorously evaluated and refined.

The estimation of EI has most frequently been achieved through self-report measures, which can be biased by misreporting (EI is typically underreported) (Stubbs et al., 2014). Unfortunately, predictors of the biases inherent to self-report data remain elusive (Rasmussen et al., 2007), so correcting for misreporting bias is currently not possible. Despite this, the accessibility and relatively low cost of self-report measures mean they are a popular choice in large scale epidemiological studies. A similarly scalable method is the mathematical modelling of EI. A notable and validated (Sanghvi et al., 2015) EI model is the NIDDK model, which is grounded in thermodynamic principles and empirical physiological data published over many decades (Hall & Chow, 2011) (for further details, see **sections 1.4.2**

and **3.5.1**). Modelling EI mathematically requires basic demographic information and the bodyweight of the subject at regular intervals, which makes it inexpensive and allows measures to be scaled up to large epidemiological studies. Indeed, the validated NIDDK EI model has been applied in a variety of contexts and populations (Göbel et al., 2014; Guo et al., 2019; Polidori et al., 2016).

Mathematical modelling shares some benefits with self-report methods (i.e. ease of dissemination and low cost) but in a critical distinction, the NIDDK model shows high predictive accuracy. Indeed, relative to gold-standard measures (i.e. intake balance method with DLW and DEXA) a negligible average deviation of ~40 kcal/day has been observed, although wide limits of agreement indicate limited precision at the individual level (Sanghvi et al., 2015). Two factors are likely to refine the models further: First, the confidence interval associated with model estimates can be minimised with more frequent bodyweight measurements (Sanghvi et al., 2015). Second, quantitative and continuous estimates of PAEE will overcome the limitations associated with assuming PAEE is constant within and between subjects (Foright et al., 2018; Sanghvi et al., 2015). Both factors can be overcome with the recent developments in digital tracking technologies.

### 9.1.1. Chapter aims

Mathematical modelling of energy balance behaviours over time facilitates novel investigations into the factors contributing to longitudinal weight outcomes after weight loss. Until now, applications of the NIDDK mathematical model have used sparse body weight measures and have lacked objective measures of PAEE. This study utilised a validated mathematical model and for the first time, integrates objective estimates of PAEE to derive changes in EI in a large sample of European adults engaged in a weight loss maintenance trial. Estimates of EI were compared amongst a series of algorithms developed and tested in **chapters 7** and **8** and amongst participants in different states of energy balance, specifically those gaining, maintaining and losing weight over 18 months.

## 9.2 Methods

### 9.2.1 Participants and bodyweight data

This chapter utilised data from the NoHoW trial. The inclusion criteria, recruitment and relevant components of the protocol have been discussed in chapter 3 of this thesis. The full experimental procedure and protocol have been published elsewhere (Scott et al., 2019). Regular bodyweight estimates were obtained from the Fitbit Aria digital scales (See **section 3.4.3.5**). To be included in this analysis, at least one weight measure every 182 days was required (which would be equivalent to one weight every clinical investigation day), although the minimum number of weights for any participant in this analysis was 33 weights in the period of observation. The body weights were smoothed by a locally weighted regression, as discussed in **section 3.4.3.5**, using a fraction value of 10%.

### 9.2.2 Energy expenditure estimation

Total daily EE was estimated using a number of approaches in this study. Firstly, in the method described in **chapter 8** using the gradient boost, neural network and random forest algorithms. One distinction from the method of **chapter 8** lay in the smoothing of estimated resting heart rate, the fraction value was decreased to 5% because of the larger datasets used in this study compared to the TEED study, and therefore the proportion of data used in the LOESS regression needed to be reduced. Estimates were then scaled to account for missingness within a day by the method described in **chapter 6**. As described in **chapter 6**, at least 18 hours per day after hourly scaling were required to be included. Any day in which no minutes could be predicted, or the sleep detection algorithm reported >20 hours of sleep were considered to be anomalous and were excluded. Furthermore, some days within the FB manufacturer estimates had extremely high PAEE values, therefore any day with PAEE > 200 kcal/kg/d or < 0 kcal/kg/d was considered to be physiologically implausible. These values would imply that the subject did not move at all in a day, or they spent every minute of the day performing vigorous activity. For example, 200 kcal/kg/day would be 15,000 kcal/day of PAEE for a 75 kg person, if this person had an RMR of 1800 kcal/day this would equate to 9-10 METs (EE/RMR) on average before any additional thermogenic costs are considered. This is roughly equivalent to running at 11– 14km/h for 24 hours continuously (Ainsworth et al., 2011). Further details on maximal time vs EE relationships is shown in **chapter 1**, figure 1.2.

Next, values were scaled to account for the EE associated with digestion, by dividing the daily estimates by 0.9. As RMR was not measured in the NoHoW trial, RMR was estimated continuously (as new weight data was observed) with the equation of Mifflin-St Jeor, which is effective in nonobese and obese individuals when compared with other common prediction equations, specifically it has been shown to more frequently fall within 10% of measured RMR (Frankenfield et al., 2005). Secondly, TDEE obtained from the FB manufacturer estimates is reported. This was calculated as the sum of minute level EE for each day and subject. Wear time was calculated as the sum of minutes where a heart rate value was measured (implying attachment on the wrist) and a valid day was one in which 1080 minutes (18 hours) were available, as per the inclusion criteria determined in chapter 6. As the FB manufacturer algorithm is proprietary, no additional adjustments for wear time or DIT were made. To be included in the analysis at least 6 valid days for TDEE data were required over the period of study. The first valid TDEE observation must have occurred in the first 12 days and a complete observation after day 360 must have been available, thereby ensuring that the participant had some data in the last third of the study. Lastly, the participants that were included in this analysis are those that could be included according to the requirements of both the FB and the hierarchical approaches, which allowed for paired comparisons to be made.

### 9.2.3 Modelling energy intake

After the above steps, EI was estimated by the mathematical model described in **section 3.5.1**, which was implemented in a Java application developed by researchers at the NIDDK. A time interval of 28 days was used to estimate $\Delta EI$, as justified in **section 3.5.1**. For the machine learning models PAEE was calculated as PAEE = ((TDEE x 0.9) – RMR) /Weight) and for the FB, PAEE = (TDEE – RMR) /Weight. It remains unclear whether DIT adjustments are made by the FB and thus estimates were not altered.

### 9.2.4 Statistical analyses

All participants were assigned to one of three groups dependant on their 18-month percentage weight changes, that is the percentage change in weight from their first weight until their last weight. Outcomes were: i) weight losers (WL) < -3% weight change, ii) weight maintainers (WLM) ≥-3% to ≤3% weight change and iii) weight gainers (WG), >3% weight change from the bodyweight at the start of the trial. The rationale for a 3% cut-off is based on a previous recommendation that the definition of weight maintenance in adults is a weight change of <3% (Stevens et al., 2006). Data were visually

represented as percentage weight change from the subject's baseline weight. At each time point, paired t-tests tested for statistical differences between models. To test for statistical differences between weight outcomes, Kruskal-Wallis rank-sum tests were used and post-hoc comparisons were conducted with Dunn's test, using the 'FSA' package in R. Where comparisons are made at numerous time-intervals, p-values were adjusted by the Bonferroni correction for multiple comparisons and significance was accepted at $p<0.05$. All statistical comparisons were conducted in R 4.0.0.

## 9.3 Results

The descriptive characteristics for the entire sample, as well as the three weight outcome groups, are shown in table 9.1. Averaged across all participants, 248 weight measures were available, 376 valid days for the machine learning algorithms and 358 valid days of FB manufacturer data were available.

**Table 9.1** Descriptive characteristics of the included sample.

Weight measures describe the number of weight measures included in the analysis, algorithm days describes the number of days available for the machine learning algorithms and FB days describes the number of days available for the FB manufacturer estimates.

| | n (Female) | Age (Years) | Height (cm) | Weight (kg) | BMI (kg/m2) | Weight measures | Algorithm days | FB days |
|---|---|---|---|---|---|---|---|---|
| **All** | 786 (562) | 45.59 ± 11.73 | 168.66 ± 8.7 | 84.03 ± 16.5 | 29.49 ± 5.16 | 248.28 ± 134 | 376.3 ± 146.15 | 358.31 ± 149.83 |
| **F** | 562 (562) | 46.57 ± 11.91 | 165.15 ± 6.66 | 80.85 ± 15.61 | 29.63 ± 5.45 | 249.64 ± 132.73 | 387.66 ± 145.21 | 369.8 ± 149.47 |
| **M** | 224 (0) | 43.13 ± 10.9 | 177.47 ± 6.77 | 92 ± 15.99 | 29.15 ± 4.35 | 244.86 ± 137.4 | 347.79 ± 144.91 | 329.47 ± 147.15 |
| **WG** | 325 (242) | 44.61 ± 11.32 | 168.76 ± 8.51 | 82.63 ± 15.67 | 28.95 ± 4.72 | 226.84 ± 127.98 | 367.77 ± 147.7 | 349.89 ± 151.15 |
| **WL** | 171 (131) | 46.26 ± 11.54 | 167.34 ± 9.02 | 86.98 ± 17.25 | 31.06 ± 5.86 | 276.1 ± 141.99 | 379.59 ± 163.44 | 361.9 ± 166.41 |
| **WLM** | 290 (189) | 46.29 ± 12.24 | 169.33 ± 8.66 | 83.87 ± 16.78 | 29.17 ± 5.04 | 255.91 ± 132.36 | 383.92 ± 133.05 | 365.62 ± 137.59 |

### 9.3.1 Weight outcomes

The distribution of weight outcomes for the groups is shown in figure 9.1 (A). At 18 months, the mean weight change overall was 1.68% and the data appear to be approximately Gaussian, with an SD of 7.85%. The observed changes in bodyweight from the start of the trial to the last weight available for each subject ranged from 31.8% gain to one 33.1% loss. Figure 9.1 (B) shows the trajectories of the groups over the 18-month observation period. In the WL group, weight loss was most rapid at the start of the observation period, with weight change from baseline of -6.2 % achieved by day 266 and reaching - 8.8% at day 546. The WLM group also showed evidence of an initial energy deficit with an average weight loss until day 322 (-0.1%) and

then a slight gain in weight towards the end of the observation. Lastly, the WG group show a more linear trajectory, gaining weight (as a % change from baseline) relatively constantly, reaching 8.2% by day 546.

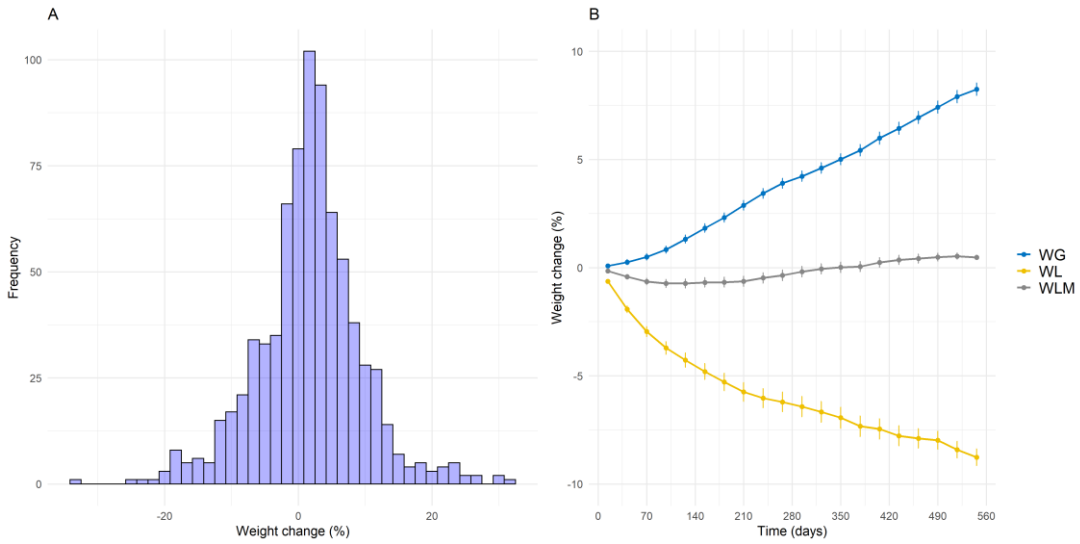



**Figure 9.1** A) a histogram detailing the distribution of weight change (%) for all included participants and B) a time series of weight change (% change from baseline), split by weight outcomes.

Weight outcomes are those losing weight (WL, n=171), maintaining weight (WLM, n=290) or gaining weight (WG, n=325), data are presented as means ± standard error.

### 9.3.2 Energy intake and expenditure changes

The change of EI between EE prediction models and weight outcome groups is shown in figure 9.2 and the variance in EI estimates at each time point is shown as boxplots and line plots in appendix 5.1. Regarding figure 9.2, it is evident that those in the WL outcome group initially changed their EI substantially, with the machine learning model predictions averaging between –413 and -411 kcal/day and the FB manufacturer estimating -427 kcal/day. At day 238, the machine learning model estimates indicated that the change in EI was positive for the WL group and the subsequent intervals tend to fluctuate close to 0. The FB manufacturer model indicated that the WL group ΔEI values were negative for all but two of the intervals examined, at day 266, where ΔEI =6 and day 462, where ΔEI =50 kcal/day.

In the first time interval for the WG group, EI increased by 43 to 44 kcal/day for the machine learning model predictions and by 26 kcal/day according to the FB manufacturer estimates. Subsequently, the remaining intervals had positive ΔEI values ranging up to 153 kcal/day (random forest, day 210).

One exception to this was at the final interval (day 546), where the ΔEI value was -27 kcal/day for the FB manufacturer. The WLM group ΔEI values in the first interval were between -90 and -89 kcal/day for the model predictions and -106 kcal/day according to the FB manufacturer. All models subsequently indicate a trend towards increasing EI in the middle intervals, for example at time point labelled day 210, the machine learning models predicted ΔEI of 86 – 95 kcal/day and the FB manufacturer predicted 54 kcal/day. Towards the last interval, ΔEI tends to decrease. The values at timepoint 546 were between 33 and 42 kcal/day for each of the models and -33 kcal/day for the FB manufacturer.

Using the EI calculated with the PAEE estimates from the gradient boost model, the Kruskal-Wallis tests and all adjusted pairwise comparisons between the ΔEI estimates for each weight outcome were significant (adjusted $p<0.05$) for all intervals prior to the interval labelled day 42, when the WLM and WG were not significantly different (adjusted $p=0.27$). The pairwise comparisons for WL vs WLM were significant (adjusted $p<0.05$) for comparisons before day 154, at which point the comparison revealed no significant difference (adjusted $p=0.075$) and WL vs WG were significantly different for all comparisons before day 238, at which point there was no significant difference (adjusted $p=0.053$).

The $\delta_0$ values (PAEE at baseline), for the machine learning models were as follows: Gradient boost = 8.01 ± 3.2 kcal/kg/day, Random forest = 8.21 ± 3.26 kcal/kg/day, Neural network = 7.1 ± 3.05 kcal/kg/day and for the FB manufacturer $\delta_0$ = 13.62 ± 4.3 kcal/kg/day. The ΔEE values are plotted for each of the groups and models in Figure 9.2. The ΔEE in the first interval (day 14) was observed to decrease in the WL group, with model predictions averaging -106 to -104 kcal/day and the FB manufacturer predicted -123 kcal/day. The WLM group also decrease their EE in the first interval, albeit to a lesser degree, with predictions averaging between -15 and -16 kcal/day for the machine learning models and the FB manufacturer predicted ΔEE = -33 kcal/day. The WG group initially increased their EE by 11-12 kcal/day according to all models but decreased according to the FB manufacturer estimates, ΔEE = -7 kcal/day.

After the first interval, each of the machine learning models predicts that the EE of the subjects in the WG and WLM group returns to slightly above their baseline for the remainder of the intervals, and the WL group trend towards 0 in the middle intervals. The FB manufacturer estimates indicate clear differences between the weight outcome groups over time, specifically, the

WL group have negative ΔEE values for the remainder of the observation period. The WLM group trend towards ΔEE = 0kcal/day, with the estimates ranging between -32 kcal/day (day 42) to 13 kcal/day (day 210). The WG group show more positive ΔEE values, which reached a peak of 63 kcal/day, at day 518.

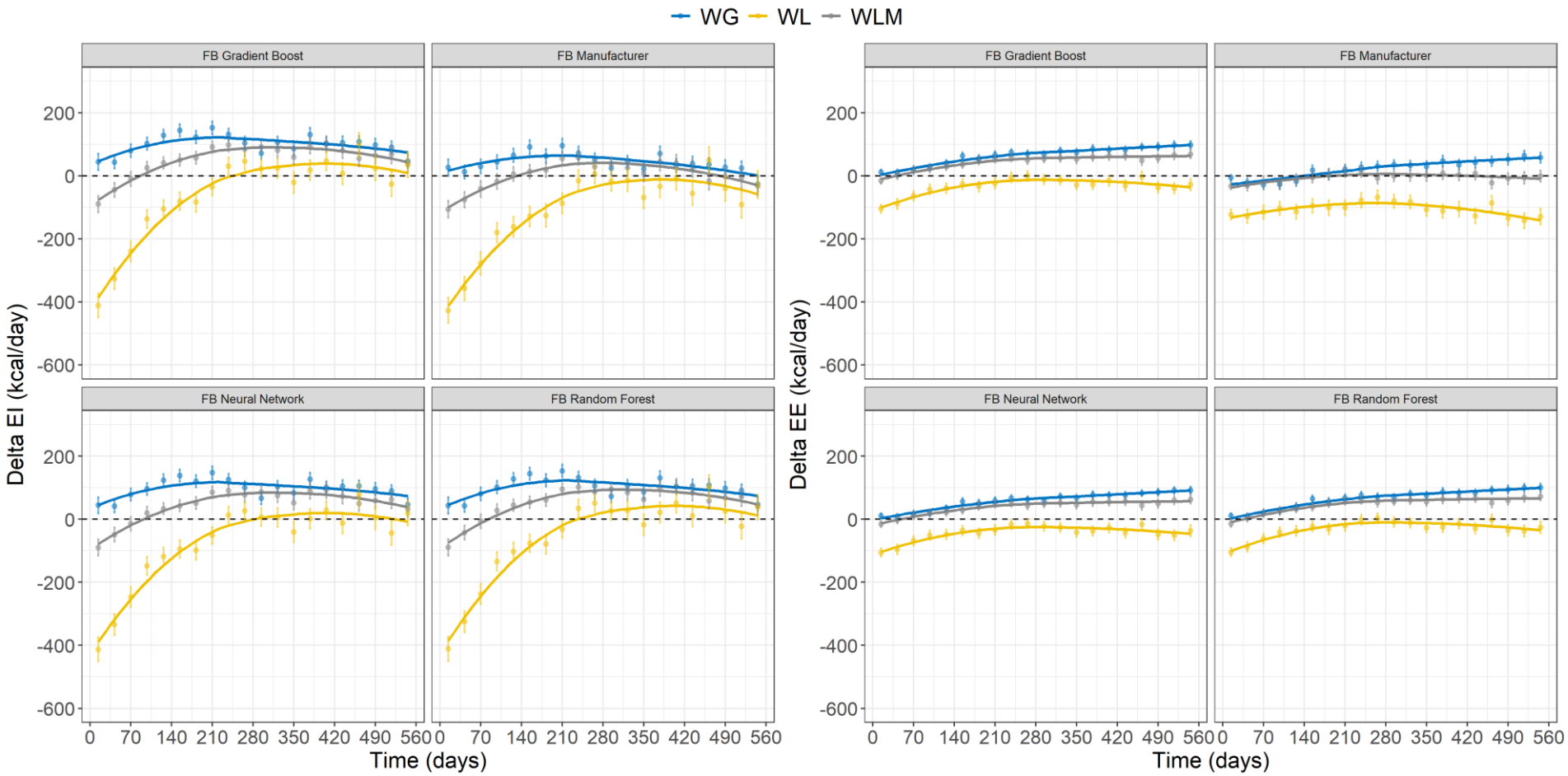



**Figure 9.2** Mean estimates of energy intake changes (kcal/day) (left panels) and energy expenditure changes(kcal/day) (right panels) from each of the models.

Weight outcomes are those losing weight (WL, n=171), maintaining weight (WLM, n=290) or gaining weight (WG, n=325), data are presented as means ± standard error.

### *9.3.3* TDEE and PAL estimates

The average estimates of TDEE and PAL are shown in figure 9.3. The PAL estimates for each model demonstrate visual evidence of a positive skew, with ranges of 1.26 – 2.21 for the gradient boost, 1.26 – 2.27 for the FB neural network, 1.26 -2.17 for the random forest and 1.22 – 2.37 for the FB manufacturer.

Estimates derived from the Gradient boost gave an average PAL value of 1.59 ±0.15 and a TDEE of 2470 ± 490 kcal/day (n=786). The neural network estimates were PAL = 1.53 ± 0.14, TDEE = 2389 ± 500 and the random forest, PAL = 1.60 ± 0.15, TDEE = 2491 ± 495. These machine learning estimates were notably smaller than the FB manufacturer estimates of PAL = 1.71 ± 0.18 and TDEE = 2663 ± 558 kcal/day (n=786), paired t-tests between each of the model predictions revealed that all TDEE and PAL estimates were significantly different ($p<0.001$).

For brevity, a single machine learning model is utilised to compare amongst weight outcomes. The Gradient Boost algorithm estimated PAL values were as follows WG = 1.57 ± 0.14, WL = 1.59 ± 0.15, WLM = 1.61 ± 0.15. The Kruskal-Wallis test revealed a significant difference amongst groups ($H = 10.74$ (2), $p = 0.005$), with WLM significantly differing from WG ($p = 0.001$) but all other comparisons were not significantly different.

The FB manufacturer comparison revealed a significant Kruskall-wallis test ($H = 6.49$ (2), $p=0.04$) and the mean PAL values were: WG = 1.69 ± 0.17, WL = 1.71 ± 0.2, WLM = 1.73 ± 0.19. A significant difference was observed between the WG and WLM groups ($p = 0.01$) but not between the other comparisons.

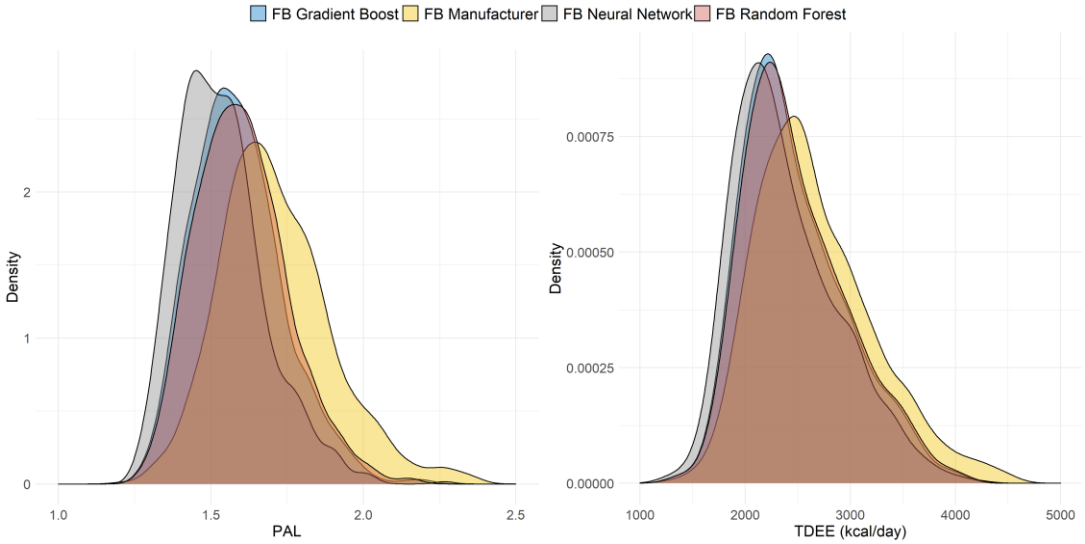



**Figure 9.3**.Density plots showing the distribution of PAL (left) and TDEE (right) for each of the methods of estimating TDEE.

## 9.4 Discussion

In a novel contribution to the fields of energy balance and weight loss maintenance, this chapter utilised digital tracking technologies to provide high-frequency estimates of PAEE and bodyweight and mathematical modelling approaches to estimate the change in EI in subjects attempting to maintain their weight. These modelling approaches demonstrated differential patterns of EI and EE depending on the long-term weight outcomes (i.e. WL, WLM or WG). It was apparent that the differences were greatest between the weight outcome groups in the first 6-months of observation. The machine learning models (i.e. neural network, random forest and gradient boost) predict that as time progressed, the EI returned to slightly above the baseline maintenance requirement (see **section 3.5.1** for further details on how this is defined) for WLM and WL groups. Similarly, the machine learning models predict that EE, which had decreased in the WL group when EI was being restricted, returns to slightly above baseline for the WLM and WG groups, and stays very close to baseline in the WL group. Each of the hierarchical approaches developed in this thesis performed similarly, but some differences were observed between the ΔEI calculated using these hierarchical approaches and the FB manufacturer estimates.

The previous literature using mathematical models consistently shows a large restriction of EI at the onset of a trial or intervention and a slow return towards baseline as time progresses (Göbel et al., 2014; Guo et al., 2019;

Polidori et al., 2016), which drives bodyweight upwards or causes a plateau in weight loss (Thomas et al., 2014). This pattern was observed in the WL and to a lesser extent, the WLM group, indicating a challenge in restricting EI over prolonged periods. From an evolutionary perspective, the invocation of physiological processes which tend to oppose weight loss and lead to regain has been selected to minimise the risk of starvation (Speakman et al., 2011; Speakman, 2007) but the extent to which these processes (i.e. increased appetite, decreased EE, etc.,) determine energy balance behaviours varies between individuals (Speakman et al., 2011) and this is apparent in these data. At baseline, participants had all been successful in achieving clinically significant weight loss, losing an average of 11.8 % in the 12 months prior to the study (Turicchi, O'Driscoll, Horgan, Duarte, Santos, et al., 2020), but the subsequent trajectories over the 18-month observation period reported in this study differ substantially. As highlighted in the introduction of this chapter, the general model of intake regulation is a model that permits a role for an array of physiological and environmental factors to influence EI variably, through interactions with genetic factors (de Castro & Plunkett, 2002). Indeed, the wide variability in EI at any time point, and the inconsistent trajectories in EI (see appendix 5.1) provide evidence for a strong role for societal and environmental factors, beyond physiological/body composition factors to influence energy balance behaviours and therefore weight outcomes. Unfortunately, understanding of the interactions between individual-level psychology, physiology, energy balance behaviours and the environment is far from complete (Stubbs et al., 2019) and the extent to which homeostatic mechanisms act on humans is contested (Müller, Geisler, Heymsfield, et al., 2018). After further validation, the objective tracking methodologies presented here could be incorporated into experiments to provide a quantitative framework of energy balance. This framework would allow other markers of human behaviour and psychology to be modelled and allow for more rigorous investigations of theories of body weight regulation.

When comparing PAL between the groups, the WLM group were observed to have the highest average PAL, although this was only statistically different from the WG group. The association between weight maintenance and PAEE has been shown previously, with those avoiding the regain of bodyweight tending to be more physically active (Catenacci et al., 2011; Jakicic et al., 2008; Ostendorf et al., 2018) and the results presented here indicate that WLM were the most physically active on average (measured by PAL), but only significantly different from the WG group. In one notable

example, conducted in a female population, Schoeller et al. report that higher PAL values (measured by DLW) after a period of weight loss were associated with less regain after a year (Schoeller et al., 1997). Similarly, Kerns et al. investigated the PAEE behaviours of contestants participating in a televised weight loss competition 6 years after the show had finished. The analysis showed that those most successful in achieving weight maintenance had a greater PAEE increase than those regaining weight (Kerns et al., 2017). One theory, termed the “energy gap” theory proposes that the reduction in TDEE from weight loss creates a “void” (i.e. change in the energy balance point needed to maintain a new reduced weight) which must be filled in order to avoid an energy surplus and weight regain (Hill et al., 2009). A higher rate of PAEE may help to fill such a void (Hill et al., 2009), thus providing a potential explanation for the associations between PAEE and WLM.

Physical activity EE may reduce in response to low energy availability (Rosenbaum & Leibel, 2010), and this offers a potential explanation for why the WL group were observed to have the lowest ΔEE in the early stages of the trial when their rate of weight loss was at its fastest. As time progressed, the machine learning models predict that the changes in EE return towards 0, whereas the FB manufacturer estimates do not and the WL group remain at values < 0. Several factors may explain this disparity. First, PAEE (which is used in the NIDDK model calculations) is calculated by taking away the predicted RMR and DIT from TDEE estimates for the machine learning models. In the FB models, no adjustment for DIT is made because it was unclear whether this is already part of the FB manufacturer algorithm. Furthermore, the Mifflin St. Jeor equation was used to estimate RMR. If the FB manufacturer incorporates a different RMR equation, which scales differently with weight changes, this step could contribute to the observed differences. Both of these factors again highlight an issue with proprietary algorithms in research studies. Also, when discussing the $\Delta\mathrm{EE}$ values, the $\delta_0$ must be considered as this value is central to $\Delta\mathrm{EE}$ calculations. The PAEE predicted by the FB manufacturer is far higher than the machine learning models (~7-8 kcal/kg/day vs 13.6 kcal/kg/day). The NIDDK models are calculated relative to a baseline requirement, which eliminates the need to estimate energy requirements and avoids the associated loss of precision (Hall & Chow, 2011). However, measures before the study were unavailable and this must be considered when interpreting these results. The fact that the machine learning models predict the ΔEE is close to 0 despite a relatively large decrease in weight in the WL group is perhaps unexpected.

Weight loss is often accompanied by changes in body composition, which should cause a decrease in EE via the loss of metabolic mass (Wang et al., 2000). It must, however, be recognised that a change in 8% of body weight for a 75kg, 175cm male, would result in a change in the RMR of ~60 kcal/day, depending on the regression equation used and assuming no other adaptive thermogenic processes. A small increase in structured exercise could feasibly offset this. It is also important to consider that the models are initialised with the average of the model predictions in the first 14 days of the study, $\delta_0$. Behaviour in this period could be related to the recent weight loss required for inclusion into the trial or through initial 'feedback' from the wearable devices.

It is important to consider the difference in the PAL values between the hierarchical approaches and the FB manufacturer in the context of typical PAL values. The lower sustainable limit for PAL in humans is considered to be in the region of 1.2, seen in non-ambulatory adults or those confined to a calorimeter without exercise. By contrast, the upper limit is thought to be approximately 4.5 seen in the most physically demanding endurance events (Shetty, 2005). The typical categorisation is as follows: 1.40–1.69 for a sedentary or light-active lifestyle; 1.70–1.99 for moderately active lifestyles; and > 2 for regular vigorous activity or a highly active job (FAO/WHO/UNU, 2004). A seminal paper by Black et al. (Black et al., 1996) collated thousands of DLW measurements and report the average PAL values for healthy, non-athlete adults. Their results showed the 40–64 years category had the lowest average PAL values of 1.69 and 1.64 for females and males, respectively. This category is the average age category of the sample analysed in the present study however, this sample had a higher BMI of ~29 kg/m$^2$ compared to ~25 kg/m$^2$ in the study of Black (Black et al., 1996). In the more recent SACN report, the distribution of PAL is reported based on two large studies (OPEN (Subar et al., 2003; Tooze et al., 2007) and Beltsville (Moshfegh et al., 2008), combined n=929) (SACN, 2011). The PAL values reported in this report are lower than that of Black (Black et al., 1996), with a mean PAL of 1.64 and a median of 1.62, ranging from 1.01 to 2.61. When these values are trimmed according to 'sustainable values' (PAL = 1.27 – 2.5), 39 subjects were lost, and the mean shifts upwards to PAL = 1.66 and the median= 1.63. The similarity to the PAL values in this study is encouraging, but further validations are undoubtedly required. Participants had a verified weight loss of approximately 11% in the NoHoW trial (Turicchi, O'Driscoll, Horgan, Duarte, Santos, et al., 2020), and it is known from data published as part of the CALERIE phase 1 trial that the caloric restriction

required to elicit such a weight reduction is associated with reductions in PAEE (Martin et al., 2011). These factors could indicate that the population in this study would have a lower PAL than previously published reference databases. When considering such reference databases it must be noted that the measured PAL values are contingent on the quality of the numerator and denominator in the PAL equation (i.e. TDEE and RMR) and the rather large assumption that the behaviours observed in the measurement period are reflective of the typical lifestyle behaviour.

### 9.4.1 Strengths

Some advantages of this work must be highlighted. This is the first time that such high-frequency body weight data has been utilised in the NIDDK mathematical model. Such regular measures, collected with digital tracking smart scales, facilitates the modelling of the variable nature of body weight (Turicchi, O'Driscoll, Horgan, Duarte, Palmeira, et al., 2020) which may be missed when less frequent measures are used (i.e. collected at clinical investigation days). Indeed, acute fluctuations in water can lead to substantial weight fluctuations (Bhutani et al., 2017) so single point or infrequent measurements are prone to error. Also, objective data from activity trackers were used and the presented data shows how widely PAEE varies in this sample. This takes steps towards minimising the biases associated with i) assuming a constant PAEE or ii) using self-report tools (Dhurandhar et al., 2015). These advantages make important inroads towards using these models on an individual level creating potential opportunities for personalised health research (Sheth et al., 2018).

### 9.4.2 Limitations

This study has some limitations to consider. Firstly, there was no gold-standard measure to compare the model predictions to. Although the plausibility of TDEE estimates has been shown in the previous chapter of this thesis, the TEED study population differ demographically from the NoHoW cohort and therefore the validity in the NoHoW dataset is not clear. Second, the mathematical model used in this study was validated in the CALERIE dataset (Sanghvi et al., 2015), which was absent of people with obesity (Rickman et al., 2011), so generalisability to the NoHoW cohort is uncertain. Next, the NIDDK model makes several assumptions about factors involved in body weight dynamics in humans, (see **section 3.5.1** for details). For example, the parameter for dietary and adaptive thermogenesis $\beta$=0.24 is likely to vary between subjects (Dulloo et al., 2012; Westerterp et al., 2004). It also does not include terms for factors such as the composition of

FFM and organ size which can influence the parameters in the model, namely the EE rate of FFM (Sparti et al., 1997). It is uncertain how these examples and other factors contribute to individual-level error, and therefore estimates of $\Delta EI$ are associated with precision estimates (Sanghvi et al., 2015). Unfortunately, such uncertainties are likely to remain because their measurement is often invasive and expensive (Dulloo et al., 2012). Lastly, the NoHoW sample may not be typical of the general population or even those engaged in an attempt to maintain lost weight. It has been suggested that as little as 20% of subjects who are overweight and then lose weight are successful in maintaining their weight loss (Wing & Phelan, 2005) and those that do may differ in psychological and behavioural characteristics (Varkevisser et al., 2019). Given that more than half of the group in this study were not classed as regainers, caution must be exercised when generalising these results to other populations.

## 9.5 Conclusion

This study combined high-frequency body weight and objectively tracked PAEE data to derive estimates of changes in EI in adults engaged in a WLM study. The modelling approaches showed variable time courses depending on the weight outcome (i.e. WL, WLM or WG). Importantly, after an initial reduction in EI in the maintenance group, evidence of increasing EI over time was observed in those losing or maintaining weight which could indicate compensatory increases in appetite in response to prior weight loss. Wide variation exists in modelled $\Delta EI$ between and within-subjects and the determinants of this variation should be examined in future studies. Differences were observed between the proprietary FB manufacturer estimates and the estimates derived through the modelling approach developed in this thesis. Gold-standard validation approaches are needed to determine which of these two offers the most precise and accurate means of deriving ΔEI and ΔEE and whether these are sufficient for individual-level research.

# Chapter 10 – General discussion

## 10.1 Summary of PhD findings

This PhD had the overarching objective of advancing the quantification of the components of the energy balance equation in free-living subjects. Energy balance in humans obeys the first law of thermodynamics and therefore, the rate of change in energy stored within the body is equal to the difference between the rates of EI and EE, per some unit of time. Given the inviolability of this law, it is possible to 'solve' the energy balance equation with an accurate and precise estimate of any two of the three components (i.e. ES, EE and EI). Obtaining accurate and precise estimates of these components is challenging, and as yet, a framework does not exist to achieve this in large, free-living studies. The approach taken in this thesis was to utilise advanced statistical approaches to learn patterns in movement and physiological data in an attempt to improve estimates of EE. Estimating EE with wearable devices has garnered much interest in health research because of the affordability, ease of use and storage capacity of many devices, but their inaccuracy represents a major barrier to their use in research settings. A mathematical modelling approach developed by researchers at the NIDDK offers an inexpensive and potentially accurate solution for modelling the dynamics of ΔES in humans, which can be linearised and solved for ΔEI. A major limitation of this approach to date has been the assumption that the PAEE of subjects is constant, in the absence of objective estimates. This is a limitation that can be overcome with wearable devices if PAEE estimates are sufficiently accurate. Therefore, accurate estimates of EE are not just important in and of themselves but can be incorporated into mathematical models to refine ΔEI estimates. Through a series of experiments, significant steps were taken towards modelling EE and subsequently EI and these are further discussed in this chapter.

Speaking in a lecture, Lord Kelvin once said*:*

*'When you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be'*.

This quote (Kelvin, 1883) captures an issue within weight loss maintenance research currently. The field is very limited in the extent to which estimates

of EI and EE can be accurately and precisely estimated during weight management interventions.

Faced with physiological resistance to weight loss and an obesogenic environment, the status quo is to regain weight after weight loss (Wing & Phelan, 2005). This is despite numerous evidence-based behaviour change interventions and public health policies. Regaining weight fundamentally implies that a subject's rate of EI exceeds their rate of EE over some time. Behaviour change interventions aiming to assist in weight management seek to influence a subject's energy balance behaviours, which are likely to be subconscious and potentially undetectable with self-report tools (Bargh & Chartrand, 1999; Stubbs et al., 2019). Thus, a situation currently exists where behavioural scientists are unable to quantify the very outcome they are seeking to influence. The ineffectiveness of weight management interventions in the long term is most likely related to the inability to precisely and accurately measure energy balance behaviours, continuously and over long durations. It is likely that should such methodologies emerge, a quantitative framework for behavioural interventions could be recognised and significant steps could be taken towards improving models and elucidating the mechanisms of action of weight management interventions. This could be achieved by linking mediators of behaviour change to objectively quantified changes in energy balance behaviours and subsequent weight outcomes (Stubbs et al., 2021).

This thesis was motivated by the need to advance measurement capabilities within the field of energy balance and the following aims were conceived:

- Investigate the validity of current wearable tracking technologies for the estimation of heart rate and EE
- Investigate methods to impute or address missing data in commercial activity monitors
- Development and validation of machine learning algorithms to predict EE, which are validated in laboratory and free-living settings
- Quantification of EE and EI in free-living subjects participating in the NoHoW trial

The following paragraphs consider these aims and the research conducted to address them. The potential implications and assumptions of this work are then presented in the context of the literature, with recommendations for future research. Lastly, the limitations of the thesis are highlighted.

### 10.1.1 Aim 1: Investigate the validity of current wearable tracking technologies for the estimation of heart rate and EE

The principles of energy balance and the recent development in physical activity trackers were discussed in the introduction (**Chapter 1**). If EE can be accurately estimated with these devices, then they can be used in free-living studies to overcome the limitations of self-report measures. **Chapter 4**, (aim number 1), reviewed the available wrist or arm-worn devices and their validity for the estimation of EE relative to criterion measures (i.e. DLW, indirect calorimeters and metabolic chambers). The trend towards utilising the wrist as a measurement site in commercial devices (Wright et al., 2017) was the motivation behind this inclusion criterion. A systematic search of scientific databases revealed 109 comparisons between different devices and a criterion. Using a random-effects meta-analysis, it was shown that the aggregate estimate of all devices was a small, but significant underestimation of EE (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020). Moderator analyses demonstrated that the inclusion of heart rate sensors within a device reduced the error in most activities, particularly cycling. Indeed, this was hypothesised because of the well-evidenced association between EE and heart rate (Ceesay et al., 1989). Moderation analyses were also conducted to investigate whether newer, commercially available devices were as accurate as research-grade monitors. No significant difference was observed overall, but commercial devices were statistically superior in ambulation and during sedentary/household tasks whereas TDEE was better approximated by research-grade monitors. There was considerable heterogeneity within devices i.e., the errors within a device differed between experimental studies with different participants, activities and protocols, in addition to the large biases associated with many devices. If a device is to be utilised in energy balance research to estimate TDEE it needs to be able to provide reasonably accurate estimates across a range of activity modalities, especially those that have been traditionally challenging for accelerometry based devices to estimate (i.e. cycling). Here, the term 'reasonably accurate' is deliberately ambiguous and what is considered sufficiently accurate will vary depending on funding constraints and the aims of the study. For instance, an energy balance study applying intake-balance methodology may prioritise the most accurate estimates of EE (at a higher economic and computational cost), and therefore minimise uncertainty in EI estimates.

A significant limitation of this work relates to the academic research cycle and the development and release of new products from commercial providers. Many of the devices included in the analysis had been discontinued and newer editions had been released. This is true of the FB device used in the NoHoW study, and only one validation study of this device could be included in the meta-analysis. The characteristics of the FB have been introduced in **chapter 3** and a critical distinction between this model compared to some previous models was the inclusion of a heart rate sensor, which allowed continuous monitoring of this physiological variable at the minute-level. The results of the meta-analysis implied that this could contribute to an increase in accuracy in EE estimates. Although the potential for accelerometer and heart rate combination approaches to improve estimates of EE has been recognised in previous devices (Strath et al., 2005), it was uncertain whether this would be the case in the FB.

In **chapter 5**, an experimental study was presented, which aimed to overcome the research question left open by the meta-analysis, specifically, how valid is the FB for estimating EE and heart rate. This study evaluated the FB, relative to criterion measures (indirect calorimetry for EE and Polar heart rate strap for heart rate). The accuracy of the FB was poor for EE in the vast majority of activity types (with a MAPE value of 44% overall) and it was only statistically equivalent to the criterion in one activity, namely running at a 5% incline (mean underestimation by the FB was ~0.5 kcal/min) (O'Driscoll, Turicchi, Hopkins, et al., 2020). In household tasks, MAPE values as high as 93% were observed. The formal definition of MAPE is presented in **chapter 3** but for additional interpretation, this would be approximately equivalent to a mean prediction of 9.3 kcal/min (assuming a constant direction of error), where the true EE was 5 kcal/min. The heart rate estimates were in far greater agreement with the respective comparator than the EE estimates, which aligned with the results of previous studies (Shcherbina et al., 2017; Wallen et al., 2016).

If the FB is utilising linear modelling to estimate EE, the direction of errors would be as observed. A characteristic of household tasks (e.g. sweeping, wiping, folding etc.) is the high velocity of movement but the low energetic cost (Ellis et al., 2016). This is different from cycling, where the movement of the wrist is minimal but the energetic cost can be large. In many activities, acceleration at the wrist is positively correlated with EE (Hills et al., 2014), and these two examples would be deviations from this general association. A small body of literature has used non-linear, machine learning approaches

to model EE in research-grade devices (Ellis et al., 2014; Montoye et al., 2015; Staudenmayer et al., 2009). These studies highlight the potential for machine learning algorithms to 'learn' the complex, activity-specific functions mapping movement, subject characteristics and physiological variables to EE, but these approaches were yet to be applied within commercial devices.

### 10.1.2 Aim 2: Investigate methods to impute missing data in commercial activity monitors

When quantifying TDEE in free-living subjects over periods longer than the battery life of activity monitors, missing data is an inevitability. Devices are removed for recharging, for water-based activities and other reasons often unknown to researchers. Thus, even if the most accurate and precise activity monitor is used, the occurrence of intermittent periods of missingness can bias activity estimates (Catellier et al., 2005) and even study conclusions (Borghese et al., 2019). This raises an important question: how might these biases be minimised? One strategy could be to permit each subject a small number of non-wear minutes and make no attempt to impute these gaps. To illustrate the problem with this approach, consider a subject with a RMR of 1800 kcal/d, who removed a device for just two hours for charging. During this time, they average ~2 METs through sedentary behaviours and office activity. This would correspond to >300 kcal expended during the period of removal although they would have 92% wear time, which falls well within the minimum requirements of most research studies. This error would be far larger if any physical activity was performed in this period of missingness or the duration of removal was larger. If the same subject performed moderate activity during this period (~4 METs), the underestimate in EE would grow to > 600 kcal. Means of addressing missingness are needed when dealing with free-living accelerometer datasets.

Research exists investigating imputation strategies in research-grade accelerometry data in short-term studies (Lee & Gill, 2018; Katapally & Muhajarine, 2014; Lee, 2013) but no research had been published relating to the commercial activity monitor data collected in long-term studies such as the NoHoW study. Therefore, it was necessary to conduct experiments to determine the optimal imputation strategy, minimum number of hours, days or weeks for valid measurements. The analysis presented in **chapter 6** (O'Driscoll, Turicchi, Duarte, et al., 2020) was conducted for these reasons and to address aim number 2. To complete this analysis a subsample of the most adherent participants in the NoHoW study (>97.5 % data availability) was used, providing over 2 million data points. Intra-class correlation

analyses provided insight into the minimum amount of data required to derive estimates of activity and the results of this analysis indicated the number of minutes, hours and days required to meet a predefined threshold of agreement (ICC = 0.9). Next, based on autocorrelation analyses it was determined that temporally proximate datapoints were far more informative in terms of imputation. With this knowledge, a simple scaling algorithm was proposed (NoHoW algorithm, see **chapter 6** for the algorithm) with the aims of a) minimising the biases of missing data and b) being computationally feasible to use in large datasets such as the NoHoW study, which has billons of minutes of accelerometer data. The algorithm was evaluated in a simulation experiment, alongside numerous other imputation algorithms varying in complexity. By holding back the 'complete' data for all subjects, then deleting data at random and using the algorithms to impute these simulated missing data, it was possible to test the bias associated with each algorithm. Secondly, a more detailed simulation study was conducted to compare the most effective individual-centred algorithms (NoHoW, Kalman imputation and multiple imputation). In this analysis, over 21,800 simulations were performed allowing rigorous investigations of the validity of the methods at varying proportions of missing data. For TDEE, comparable errors were observed between the NoHoW algorithm and multiple imputation above ~16% missingness, with maximal RMSE values of ~69 kcal/day in the simulation with the largest proportion of missing data. Critically, the computation time of multiple imputation was over 450 times greater than the NoHoW algorithm (17 minutes vs 2.1 seconds per iteration). In attempting to apply this to a dataset the size of the NoHoW study, exponential growth in run time may be observed, which would make this approach infeasible with current computing capabilities. In addition to being a novel methodology in this field, there are important implications beyond EE estimation, for example, it was shown to be an accurate imputation method for both steps and physical activity categories. The subsequent chapters consider methods to model EE from wearable devices. However, if the devices are not worn, there is no data on which EE can be modelled and even the most accurate model would be of limited use. The clear implication of this work is a simple and accessible method, which offers a means to offset a potentially large contributor to errors in physical activity and TDEE summaries collected from a FB.

### 10.1.3 Aim 3: Development and validation of machine learning algorithms to predict EE

The main findings from **chapters 4** and **5**, and aim 1, were that commercial activity monitors such as the FB provide inadequate estimates of EE. The extent to which researchers can address these issues in commercial devices is limited because of the proprietary nature of the prediction algorithms. Despite the FB's accuracy limitations, several important characteristics of this device make it a scalable solution (in terms of duration and the number of participants) for real-world medical and health research. The devices are economically viable at ~£100 per device, they are durable and have cloud storage capabilities (Rosenberg et al., 2016; Vooijs et al., 2014; Wright et al., 2017). It was known that complex, non-parametric learning algorithms i.e. tree-based methods (Ellis et al., 2014) and artificial neural networks (Montoye, Begum, et al., 2017; Staudenmayer et al., 2009) can be used to model accelerometer data but whether these advanced techniques could be used to improve estimates of EE in commercial devices was uncertain. This work was the first attempt at such approaches in commercial devices (O'Driscoll, Turicchi, Hopkins, et al., 2020).

To develop models in a supervised setting, predictor variables are required as well as a gold-standard outcome variable, which is to be predicted based on the time-matched inputs. At present, gold-standard EE data can only be provided by indirect calorimetry at the required epochs (1 minute or less) and thus, a pair of laboratory studies (described in **chapter 3**) were combined in **chapter 7**. Both studies provided accelerometer data, as well as numerous physiological inputs from 3 different device configurations. In the combined protocols which consisted of sedentary, household and exercise tasks, the predictive accuracy of the algorithms used (random forests, gradient boosting or deep neural networks) for EE and activity intensity exceeded that of one of the more accurate research-grade devices, the SWA (O'Driscoll, Turicchi, Beaulieu, Scott, et al., 2020). In the generalisability studies, when applied to out-of-sample datasets, some degradation of performance was observed although these models still tended to produce RMSE values below the SWA manufacturer estimates, despite the reduced training data and notable differences in protocols and participants. This observation raises important considerations; it may be the validation method used (LOSO) provides an overestimate of the accuracy that could reasonably be expected in unseen participants. Alternatively, the degradation of performance may relate to the differences between the

activities performed in the protocols, and generalisability would be improved with more activities contributing to the training set.

Whilst the model development and training phase needed to be in a controlled laboratory setting, the true test of these models would be in a more ecologically valid, free-living environment, where the type and duration of the behaviour of the subject are not controlled or known. Indeed, the development of obesity occurs in a free-living environment and this illustrates the importance of more ecologically valid validation studies. A 14-day study was conducted in which 30 subjects wore several wearable devices to provide minute-level movement and heart rate data. The algorithms presented in **chapter 7** were utilised to derive minute-level estimates of EE and through a hierarchical modelling approach, TDEE. This study was the first time machine learning models have been evaluated in this manner and will be the first comparison with the gold-standard (DLW), the analysis of which has been unfortunately delayed due to the disruption caused by the ongoing COVID-19 pandemic. The SWA was utilised as the next best solution, though this is not considered to be a gold-standard comparator.

The hierarchical models tended to overestimate TDEE relative to the SWA manufacturer estimates, which was also seen in the PAL and EI estimates, as TDEE is central to the calculation of these outcomes. This chapter probed the potential causes of this disparity and highlighted some important points which may aid in gauging the plausibility of these estimates. First, the SWA has known limitations in athletic populations (Koehler & Drenowatz, 2017), which many of the subjects tested in this chapter were likely to have been (as indirectly indicated by a relatively high step count and a relatively low body fat percentage). This is potentially attributable to the use of the WHO equation in the manufacturer models. This widely used RMR prediction equation was derived on 7000 individuals from 23 countries but its accuracy has been called into question previously (Müller et al., 2004). In a sample with a higher than average percentage FFM for their body weight, it is expected that RMR would be higher than predicted by a linear model based on body weight (Schofield et al., 2019). Indeed, the WHO prediction equations systematically overestimate RMR at low RMR values but underestimate RMR at high RMR values according to a previous study (Müller et al., 2004), and the plots in figure 8.8 of this thesis indicate this. The effect of utilising predicted rather than measured RMR was directly investigated by running the hierarchical models with the WHO equations and

agreement improved substantially, providing strong evidence that the RMR equations may explain the substantial proportions of the difference between the SWA and machine learning model predictions.

Second, comparisons were made to a similar study where DLW data were reported (Shook et al., 2018). It was argued that as the average RMR and step count in the TEED study were both notably higher, it would be reasonable to assume that the TDEE values (measured by DLW) would be higher than the 3170 kcal/day value reported in the highest TDEE group (Shook et al., 2018). Nonetheless, DLW data are required to confirm these suppositions. A subsequent and important distinction here between the manufacturer estimates (SWA and FB) and the modelling approaches is the transparency of the analysis. When the method of producing estimates of EE is completely proprietary, researchers applying these models have no understanding of the 'under the hood' assumptions and therefore cannot ascertain whether these models are appropriate for their experimental settings. Whilst the specific algorithms used in the models presented in this thesis are highly technical, the assumptions of the models, development cohorts and data are transparent which will aid interpretation to an extent.

### 10.1.4 Aim 4: Estimation of EE, EI and energy balance in the NoHoW trial

A central concern in the literature related to this thesis was that EI and physical activity data are typically collected with self-report questionnaires and these are of limited validity. It has been argued that they *'no longer have a justifiable place in scientific research aimed at understanding actual EI and actual PAEE'* (Dhurandhar et al., 2015. pp 2). Quantitative mathematical approaches for estimating EI do exist and have been reviewed in **chapter 1**, however, the lack of an objective PAEE measurement almost certainly contributes to the limited precision observed at the individual level (Sanghvi et al., 2015). If TDEE could be estimated continuously and accurately it would almost certainly refine estimates of EI using such models (Sanghvi et al., 2015).

The aims and studies discussed above may be considered as prerequisites for achieving the final aim of this thesis: to model the components of energy balance in the participants of the NoHoW study over weeks, months and years. By incorporating different means of estimating TDEE in the mathematical models (described in **chapter 3** and **9**), change in EI could be estimated and compared between groups with different weight outcomes as well as between predictive models. In applying these modelling approaches

to the NoHoW dataset, it became clear that patterns of EI vary over the time course of the study, particularly in the WL and WLM groups. This trajectory in $\Delta EI$ was remarkably similar to previous modelling studies (Göbel et al., 2014; Guo et al., 2019; Polidori et al., 2016). The results suggested that EI was restricted to the greatest degree in the WL group. The EE also fell and gradually increased towards and above baseline as the restriction subsided, although the FB manufacturer estimates did not show such a pattern. The changes in the group averages over time were slight, but the variability in PAEE (expressed in PAL) at the individual-level confirms that these estimates are important to include in these modelling studies. Future modelling work will aim to gain a more quantitative understanding of the importance of high-frequency body weight measures and PAEE, with particular consideration to the effect of having varying degrees of data availability. Nonetheless, the change in PAEE was relatively minor when compared to that of the EI changes, which is exactly as suggested in previous mathematical modelling studies (Polidori et al., 2016).

## 10.2 Implications of this work and areas of future research

The estimation of TDEE in energy balance research to date has involved a trade-off. If accuracy is the primary objective then the DLW method can be used to obtain the average TDEE over a period of ~14 days. The price of this method in addition to participant and researcher burden means that small groups of subjects must be studied and repeated measures are often not possible. Research-grade accelerometers may also be used, but the requirement for recharging means that they are also limited to relatively short-term measures and many lack the required precision and accuracy. Lastly, if scale and longitudinal measures are required then self-report tools can be used, but they also have substantial accuracy issues. With the recent developments in consumer wearable devices and the *internet of things*, it is now possible to overcome the limitations of scale and so-called ‘snapshot’ measures with relative ease, paving a new path for research in this area. What is broadly lacking, however, is an understanding of how these devices can be integrated into energy balance research. This thesis has taken important steps towards progressing wearables research within the field of energy balance by combining novel approaches and adaptations of previous mathematical modelling studies. The subsequent paragraphs consider the potential applications within the field of energy balance and related health research fields.

One of the most important benefits of being able to quantify EE continuously is the potential to further understanding of the relationships between PAEE and long-term health outcomes. It is important to state here the distinction between physical activity and EE; The former is defined as any bodily movement that results in EE (Caspersen et al., 1985) and thus EE increases with physical activity (Hills et al., 2014). It is known that physical activity (resulting in EE) and the minimisation of physical inactivity can positively impact non-communicable disease and mortality risk (Lee et al., 2012). However, many of the studies contributing to this evidence recruit large samples and follow these samples over many years, and have therefore had no other option but to assess sedentary and activity behaviours with self-report measures (Physical Activity Guidelines Advisory Committee, 2018). As with dietary self-report measures, these estimates are often misreported although the bias tends to be an overestimate and this is related to social desirability biases (Adams et al., 2005). This is problematic because mis-reported measures distort the observed relationships between biomarkers of health and disease and activity behaviours. For example, Celis-Morales et al., report that regression coefficients for the relationship between MVPA and HOMA-IR, insulin and triglyceride are up to 50% lower when using self-reported physical activity (IPAQ), rather than accelerometer-derived estimates (Celis-Morales et al., 2012). Of the relatively rare studies that do objectively estimate physical activity with accelerometers, the vast majority use research-grade devices (Dohrn et al., 2018; LaMonte et al., 2018; Lee et al., 2018), and are therefore restrained to very short measurement periods. In a recent meta-analysis, long-term health outcomes are inferred based on measurement periods meeting the following criteria: *'We included all participants who recorded a wear time of 10 or more hours each day for four or more days'* (Ekelund et al., 2019). Assessing PAEE by the methods presented in this thesis gives a clear opportunity to extend these observation periods over weeks, months and years, whilst still retaining the capability to assess EE at the minute-level.

Currently, reference DLW databases are extremely expensive and resource-intensive to obtain. Indeed, the international atomic energy agency (IAEA) DLW database is a relatively recent venture which has collated DLW measurements published since 1981, and at the time of writing this stands at 7479 subjects from 361 studies and 31 countries (IAEA DLW database, 2021). Given the time and resource invested in these studies, this is a relatively small sample compared with what can be achieved with activity monitoring devices, which can be used to track the behaviours of hundreds

of thousands of subjects (Doherty et al., 2017). Indeed, there were an estimated 29.57 million Fitbit users between 2012 and 2019 (*Statista*, 2021). By bridging the gap between these devices and the quantification of EE, as this thesis has attempted to do, it is likely to be possible to increase the size and information contained within such datasets by many orders of magnitude. A natural extension of TDEE estimates is the estimation of dietary reference values or EI requirements of groups of subjects (SACN, 2011). This is a critical area of research that currently has large gaps in the available evidence, specifically for adults aged 18-30 years and >80 years (SACN, 2011). The ease of measurement with tracking technologies may fill this gap and allow large datasets to be collected in previously understudied groups. Moreover, DLW data offer limited insight into hourly, daily, weekly or seasonal EE in subjects. There is variation in participation in physical activity depending on climate and day of the week (Aspvik et al., 2018; Chan et al., 2006; Doherty et al., 2017; Merchant et al., 2007; Shiroma et al., 2019) and it may be feasible to add these dimensions into reference databases in the future. The above is conditioned on repeated and robust evaluations of the accuracy of tracking technologies for EE estimates in a range of populations.

Longitudinal tracking of EE and EI has implications for fields related to energy balance. The knowledge of a subject's energy balance over time is a critical but often overlooked component in health research. Both acute and longer-term fluctuations in energy balance are implicated in blood lipid and glucose dynamics (Frayn & Evans, 2019. pp 277 - 301) but estimates of energy balance are infrequently incorporated in such studies. For example, studies (Hjorth et al., 2017; Ritz et al., 2019) have investigated the interactions between fasting blood glucose and weight outcomes and have produced results suggesting that impaired glucose uptake leads to reduced satiety in high carbohydrate diets. Insulin and glucose metabolism are seemingly altered at different rates of energy throughput, irrespective of whether a subject is in energy balance or not (Büsing et al., 2019). High vs low energy turnover may influence sensations of hunger and hormones related to appetite (i.e. GLP-1 and Ghrelin) (Hägele et al., 2019) or EI (Beaulieu et al., 2018; Edholm et al., 1955) and energy imbalances are also known to impact glucose metabolism (Lagerpusch et al., 2012). In the absence of an energy balance framework, there is the potential for the energy balance of the subject to confound the outcomes of such metabolic research, and this limits the conclusions that can be drawn.

In a recent landmark study, Berry et al report the results of a random forest model predicting the rise in triglyceride at 6 hours postprandially (Berry et al., 2020). Despite a comprehensive set of features fed into the machine learning model, including microbiome, sleep, anthropometric, dietary, immunological and physical activity variables, the predictive ability for triglycerides was unremarkable (r=0.47 in the training cohort and r=0.42 in the validation cohort). Triglyceride responses are particularly sensitive to EE and energy balance status, as well as the activity status of the individual (Maraki & Sidossis, 2010) but in this study, no measures of energy balance were available. Minutes of physical activity were estimated by a research-grade accelerometer, but accelerometer methods alone have been demonstrated to have limited accuracy for the classification of activity (Montoye et al., 2018). It is likely that the lack of physiological variables in the prediction of activity intensity may make it challenging to distinguish between the activities with similar acceleration patterns but different energy cost (i.e. walking on a flat surface vs carrying a load or walking on an incline) (Lyden et al., 2011). Based on the methods presented in this thesis, it may be possible to incorporate important measures of energy balance behaviours in future studies of glucose dynamics, where energy balance is likely to be implicated.

A further implication of this work is the potential for advancing other scientific fields related to weight management. The field of genetics has seen substantial expansion in recent decades due to advances in sequencing technology, as well as statistical and computational techniques to process the datasets. Genome-wide association studies are exploring the genetic basis of EI, EE and associated health outcomes (Cole et al., 2020; Jiang et al., 2018). Some have argued that this methodology has not added much to our understanding, in part due to crude measures of the behavioural phenotype (Müller, Geisler, Blundell, et al., 2018). These studies require extremely large samples, which has necessitated estimates of energy or dietary intakes by self-reported measures over small time frames. This leaves little to no possibility to investigate the interactions between genetics, energy balance behaviours and chronic cardiometabolic conditions such as diabetes and heart disease (Jiang et al., 2018). Advancing such research, by incorporating objectively tracked measures of energy balance behaviours, would allow novel investigations into the determinants of weight and health outcomes, and the rigorous assessment of theories of body weight regulation which implicate genetic factors, such as the general model of

intake regulation (de Castro & Plunkett, 2002) which was discussed in **chapter 9**.

Another example would be the psychological and behavioural sciences related to weight management. It is evident from the plots presented in **chapter 9** and **appendix 5.1** that there exists a wide range of EI at any given time point and hidden within weight trajectories are periods of over and undereating on an individual level (Chow & Hall, 2014). The argument has been made that measurement of psychometric variables at fixed time points (i.e. clinical investigation days at 6-month increments) is oversimplistic because they ignore the periods between measurement days (Stubbs et al., 2019). This is problematic because the autonomous processes which are implicated in fluctuating energy balance behaviours (e.g. emotions) can change rapidly (Bargh & Chartrand, 1999; Stubbs et al., 2019). To establish a causal understanding between these factors, continuous measures of the predictor and outcome variables are necessary. Continuous ecological measures (real-time psychometric measurements in ecologically valid environments) collected with mobile applications have been used previously but it seems the methodological quality of these tools is poor and there is a lack of standardisation between studies (Degroote et al., 2020). The integration of high-quality momentary assessment tools with objective measures of energy balance can potentially contribute to a greater understanding of cause-effect relationships between psychological states and behaviour and the necessary refinement of interventions (Stubbs et al., 2019). Though emphasis must be placed on the reliability and validity of the assessment tools used in this work, a review of >450 papers showed no correspondence between EI and appetite ratings in >50% of included papers (Holt et al., 2017).

At this point, it is important to reflect on what the models presented estimate and where they can and cannot be applied. Concerning EE/activity, these measures give minute level estimates. They do not provide information on the types of activity being performed. To understand the determinants and barriers to physical activity and design effective interventions to improve this outcome, understanding the type and context of activity being performed will be important (Burton et al., 2012; Doherty et al., 2013; Koorts et al., 2011). As with EE, categorisation of activity behaviours represents a significant but not insurmountable computational challenge and significant progress has been made using wearable cameras to collect ecologically valid labelled training data (Doherty et al., 2013). Harmonisation of these approaches

would offer exciting avenues to understand how, where and when EE is accumulated in individuals or groups, which could be extended to personalised strategies to change physical activity behaviours.

Concerning EI, the precision of estimates from the NIDDK models would probably become unacceptably low at <1-2 weeks. The work presented here is entirely focussed on estimating EI from EE and physiological models, however, this is just one piece of this puzzle and these methods say nothing about the short-term pattern of EI, meal composition and macronutrient intake, which are all associated with an array of metabolic processes, behaviours and health outcomes (Beaulieu et al., 2017; Byrne et al., 2017; Holt et al., 1995). Food groups tend to be misreported also, although the specific relationship between the degree of misreporting and food groups or even the 'healthiness' of the food is uncertain (Garden et al., 2018). The solution to this problem may lie in combining the methodologies presented here with recent developments in urinary metabolic phenotyping, which is showing potential to estimate dietary patterns of subjects. A recent study conducted in nearly 2000 US adults showed that some urinary metabolites covary with the consumption of self-reported dietary nutrients, although associations are moderate ($r = 0.1 – 0.6$) (Posma et al., 2020). An alternative may lie in the use of ecological momentary assessment with smartphone apps. The 'Smart-intake application' prompts participants via e-mail before each meal, reminding subjects to take a photograph of the meals they are consuming (Martin et al., 2012). The omnipresence of mobile phones makes them a useful tool to capture eating behaviours in an ecologically valid environment. Importantly, this is still subject to participants remembering to photograph their foods and uncertainty remains regarding the accuracy of these methods. In a validation study compared to DLW, EI was 63% of the TDEE value in a sample of 23 women with obesity, although in this study, this value could be improved with the removal of some erroneous days (defined as days where reported energy intake was 1) <60% of TDEE, 2) <1000 kcal, or 3) <2 meals (not including snacks) were consumed (Most et al., 2018). A potential limitation of food photography approaches is that researcher time is likely required to process or verify photographs. Substantial progress has been made in the area of computer vision, in which deep neural networks can be used to derive the energy and macronutrient content of foods, based on a recording or static images. This approach requires algorithms to classify the type of food based on surface colours, shapes and texture and then estimate the volume of the food, both of which are associated with significant challenges, which have been

reviewed recently (Lo et al., 2020). Errors in volume estimation using state of the art methods vary, but are typically <20% (Lo et al., 2020). Classification of foods in images is slightly more advanced. Training and testing on the Food-101 database, which consists of ~100,000 food images (25% of foods are retained for testing), benchmarks currently stand at accuracies of > 95% (Foret et al., 2020). Both metabolomics and food photography approaches create substantial researcher burden and do not appear, currently, to provide the level of accuracy required. Food recognition approaches are rapidly developing but are not yet established. As these approaches continue to develop and evolve, exciting opportunities to integrate dietary intake estimates within an energy balance framework will surely arise.

## 10.3 Assumptions and considerations

Research progresses through technological advancement, scientific breakthroughs and collaboration between fields. The models presented here are more sophisticated than models such as 'flex-HR' (introduced in **chapter 1**) and other similar linear modelling approaches. This complexity appears to result in more accurate estimates of EE, at least as shown in the laboratory studies conducted in this thesis. However, as with any predictive model, they are associated with several assumptions, which may not necessarily hold for all subjects in all situations. A critical step in refining methodologies is to understand the potential consequences of the assumptions, thus, the work here must be 'stress-tested' to further understand their limitations and applications.

The case has been made throughout this thesis that there is a significant advantage to the transparency of the modelling approach taken. It must be stated that there are some inputs to the models which rely on proprietary algorithms. The movement and heart rate variables in the models are extracted from Fitbit. Fitbit utilises proprietary algorithms to convert raw acceleration and photoplethysmography signals into activity estimates and heart rate, respectively. These algorithms probably apply filtering and cleaning steps and importantly, have the potential to be altered with firmware updates (Nelson & Allen, 2019). The solution to this issue would be the open-access of the raw data from commercial companies and products, as is offered by some research-oriented companies (Bassett et al., 2012). Without access to this information, it will be important to conduct agreement studies between different firmware versions to ensure consistency in outputs. This leads to a consideration of what the 'ideal' device might look

like for research studies. Undoubtedly, cloud-connectivity, long-battery life, comfort and acceptance by participants is imperative for use in research studies, to facilitate long-term assessment. An ideal device would also provide access to raw accelerometer and physiological signal data, at sufficiently small epochs and the cleaning steps applied by the manufacturers, with any alterations made to with software updates. Lastly, it would be possible to hide self-monitoring and motivational capacities of the device and the respective apps, therefore allowing researchers to untangle the effects of interventions and that of simply wearing a device and interacting with commercial applications, which often include motivational content.

When generalising these models to free-living, it is assumed that sleeping metabolic rate is ~95% of RMR because this is the average value observed in humans. However, there is likely to be variance around this value and the sleeping metabolic rate can feasibly range between 0.85 - 1.02 x RMR (Goldberg et al., 1988), and may vary based on the composition of the diet (Lejeune et al., 2006). Next, the models are centred around RMR, which was not measured in the NoHoW trial. Some evidence suggests that as weight is lost, metabolic adaption occurs to give a metabolic rate lower than what might be expected based on the composition of the weight lost (Wolfe et al., 2018). This metabolic adaption might persist over many years and even after the regain of lost weight (Fothergill et al., 2016), though others report contrary results, with limited or no evidence of metabolic adaptation after weight has stabilised (Amatruda et al., 1993; Das et al., 2003; Wolfe et al., 2018). These contradictory results mean it is challenging to precisely quantify the effects this might be having in the NoHoW cohort without direct measurement. Prediction equations for RMR such as Harris-Benedict, WHO or Mifflin-St Jeor may provide reasonable predictive accuracy at the group level, but are often associated with large errors at the individual level, indeed, a study conducted in 30 healthy adults reported that common prediction equations had a small mean bias (-14 kcal/day) for Harris-Benedict but wide limits of agreement (> 300 kcal/d) (Flack et al., 2016). Where RMR data are unavailable, which is typical in large scale studies such as NoHoW, predicting RMR is unavoidable. Unfortunately, errors in RMR estimates may silently propagate through the models, as the denominator for METs is RMR. The Mifflin-St Jeor equation was used in **chapter 9** and it appears to be one of the most reliable equations in obese and non-obese subjects (Frankenfield et al., 2005) but any linear model is likely to have limited accuracy at the individual level. Indeed, errors of up to

~600 kcal/day were observed when comparing the WHO equation to RMR measured by the GEM in **chapter 8**. Some recent evidence suggests machine learning approaches may predict RMR more accurately in clinical populations (Ponce et al., 2020).

There will naturally be subjects that deviate from the assumptions of the NIDDK model (see **chapter 3**), even though rigorous work has been conducted to consider these issues (Hall & Chow, 2011). This is also true of the EE algorithms which will tend to reproduce the training data and the complex function within that data. This thesis combined these two modelling approaches and the potential for model errors to be compounded must be recognised. Another significant limitation in this work is the assumption of a constant rate of DIT for all subjects. The extent to which some EE attributable to digestion may be incorporated in the calibration data has received extensive consideration throughout **chapter 8,** and those analyses and considerations are not repeated here. This issue has been investigated in the development of the NIDDK model. To account for digestive and adaptive processes, the model uses a value of 0.24. This value is estimated on data collected in 8 longitudinal weight-loss studies and 157 subjects (Hall & Jordan, 2008). Importantly, Hall and Jordan report an associated standard deviation of 0.13 based on Monte-Carlo simulations, indicating a relatively large degree of uncertainty in this estimate. An alternative approach used in previous studies is to estimate DIT based on the self-reported energy and macronutrient dietary intakes (Brage et al., 2015). Utilising this approach may allow researchers to account for dietary factors which alter the DIT of a subject (i.e. high protein intake (Westerterp et al., 2004)). As with EI, macronutrients can be over or under-reported (Macdiarmid & Blundell, 1998), which could introduce substantial error in DIT estimates. It is therefore unclear whether using self-report data would offer any benefit beyond assuming a constant DIT factor amongst participants.

An alternative source of error may relate to the phenomena of exercise economy, which describes the energy cost of mechanical work, independent of body weight (Pontzer, 2017). Very generally, the energetic cost of locomotion appears to be negatively correlated with the training status of the individual (Morgan et al., 1989; Saunders et al., 2004) although this finding is not completely consistent, with other studies reporting no differences based on the distribution of body mass (Browning et al., 2006). Those classified as obese may have worse economy during locomotion (Chen et al., 2004) although again, this difference is not always observed (Browning et al.,

2013). Weight loss in combination with an exercise regime may serve to increase economy of exercise in previously sedentary adults, although this is not seen if exercise is not included in the weight loss regime (Amati et al., 2008). Weight loss may serve to decrease the metabolic cost of isometric muscular contractions (Peyrot et al., 2012) or may act on exercise economy via hormonal pathways, perhaps by a reduction in leptin concentrations and associated adrenal and thyroidal changes (MacLean et al., 2011; Rosenbaum & Leibel, 2010). Overall, it appears that characteristics such as age, weight, sex, and cardiorespiratory fitness explain some of the variation in exercise economy (Chen et al., 2004) but the mechanisms remain uncertain.

At present, factors such as DIT and exercise economy are not measurable to the degree that they may be used as input variables in prediction models. It must be recognised that all of the above may lead to errors for individual subjects and future work must aim to gain a quantitative understanding of these phenomena, to model TDEE and EI more accurately. It is of paramount importance that future research considers the above and quantitatively investigates the implications of these assumptions. This will be necessary to move towards applications in clinical populations.

## 10.4 Limitations of this PhD

In each chapter of this thesis, specific limitations have been raised for the methodology employed in that specific chapter. The aim here is to consider general limitations which apply to the body of research overall but not repeat study-specific limitations. Furthermore, the above assumptions can also be considered as limitations of the models in their current form.

First, the sample used to study energy balance and WLM in **chapter 9** may limit the generalisability of results. The sample analysed were those that i) lost a substantial amount of body weight before the trial ii) on average, maintained lost weight (though substantial individual differences exist) and iii) did not drop out of the trial. The potential that this sample is of limited representativeness must be considered. Furthermore, the sample used for development in **chapter 8** were healthy adults, for whom the relationship between the input vectors and outputs (i.e. EE or activity classification) may differ from samples with various diseases. Above, it was suggested that wearables may be used to study the energy requirements of previously understudied age groups and if this extended to various disease states, the collection of new or additional calibration data would be required.

Almost all the results and arguments made throughout are relevant to the use of the FB device. This may be considered a limitation because devices themselves could be influencing the subject's behaviour. Mobile applications provide the subject with various rewards and goals, delivered through apps, which can motivate the subject (Lyons et al., 2014). It may also mean that these methods would disproportionately apply to certain groups that are most likely to engage or be able to engage with this technology. Research has implicated gender, ethnicity and psychosocial metrics in user engagement (Lewis et al., 2020) and it may be that factors like education status, body weight and physical activity level may be predictive of decreased ownership of wearable devices (Macridis et al., 2018). It will be important to ensure representativeness across all of the aforementioned strata in future uses of these models.

A further limitation of this work relates to the methodology for the collection of body weights. The Aria scales were used (see **Chapter 3**), however, controlling each measurement is infeasible and subjects may have weighed themselves in different clothes, or at different times of the day. To overcome this, a smoothing regression was fitted to the weight data although it is currently impossible to accurately quantify the degree of noise in this data. In free-living subjects, EI can fluctuate markedly (Bray et al., 2008; Tarasuk & Beaton, 1991), which, because of the associated carbohydrate and sodium, causes fluctuations in total body water and weight (Durnin, 1961; Edholm et al., 1970; Hall & Chow, 2011). Water fluctuations play an important part in the composition of weight change. Bhutani et al used biomarker methods and DEXA to show body water fluctuates significantly with weight change and that short-term weight change was composed of 84% FFM (Bhutani et al., 2017). Such issues could bias the EE models (as weight is a predictive variable), although the simulation studies reported in **chapter 8** imply large variance in weight leads to small or no changes to the METs output. Regarding the EI models, Hall and Chow, using simulated data, illustrate the effects of water balance changes on the NIDDK model (which does not currently account for large fluctuations in body water) (Hall & Chow, 2011). Their results indicate that change in EI is likely overestimated when weight loss is rapid because this phase of weight loss is characterised by a proportionately high loss of body water (Heymsfield et al., 2011). Hall and Chow suggest that an increased frequency of body weight sampling can overcome this limitation (as linear regression is fitted to the weights), and in this sense, the data used in this thesis is unrivalled in the published literature on mathematical modelling of EI.

## 10.5 Conclusions

Some research within and related to energy balance has been in crisis because of an inability to quantify both EI and EE longitudinally. If accurate estimates of EE could be obtained from wearable devices (e.g. FB) they could be used to estimate EI and EE in large samples in free-living environments. This thesis began by investigating the validity of wearable devices, and it was shown that many are inaccurate for the estimation of EE. Subsequently, a computational approach was taken to develop and evaluate algorithms to address missing data and then to improve estimates of EE, by learning the non-linear relationships between EE, movement, and physiological variables. Taken together, the developed methodologies were applied to approximate TDEE and were incorporated into mathematical models to provide $\Delta \mathrm{EI}$ estimates in participants in the NoHoW study. These studies provided strong evidence that EI varies more than EE in subjects gaining, losing or maintaining weight. Furthermore, it was shown that in those losing or maintaining weight, an initially large change in EI slowly returns towards baseline, indicating a slow relaxation of energy restriction. Whilst this thesis illustrates the potential value and utility of such approaches, much work is required to further develop accurate, precise and accessible methods for energy balance modelling. These methods would be enhanced by cloud-connected devices that provide raw acceleration in three axes and physiological signal sensor data. If these methodologies are developed, tested and validated further, they could potentially offer a scalable solution to objectively quantify energy balance behaviours across a multitude of life sciences.

## References

Abbott, W. G. H., Howard, B. V., Christin, L., Freymond, D., Lillioja, S., Boyce, V. L., Anderson, T. E., Bogardus, C., & Ravussin, E. (1988). Short-term energy balance: Relationship with protein, carbohydrate, and fat balances. *American Journal of Physiology - Endocrinology and Metabolism*, *255*(3 (18/3)), E332–E337. https://doi.org/10.1152/ajpendo.1988.255.3.e332

Achten, J., & Jeukendrup, A. E. (2003). Heart Rate Monitoring. *Sports Medicine*, *33*(7), 517–538. https://doi.org/10.2165/00007256-200333070-00004

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Cunningham, J. E., Fulton, J., & Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports of physical activity. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwi054

Agha, M., & Agha, R. (2017). The rising prevalence of obesity. *International Journal of Surgery Oncology*, *2*(7), e17. https://doi.org/10.1097/ij9.0000000000000017

Ahmadi, M. N., Chowdhury, A., Pavey, T., & Trost, S. G. (2020). Laboratory-based and free-living algorithms for energy expenditure estimation in preschool children: A free-living evaluation. *PLoS ONE*, *15*(5), e0233229. https://doi.org/10.1371/journal.pone.0233229

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., & Leon, A. S. (2011). 2011 Compendium of Physical Activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, *43*(8), 1575–1581. https://doi.org/10.1249/MSS.0b013e31821ece12

Al-Ati, T., Preston, T., Al-Hooti, S., Al-Hamad, N., Al-Ghanim, J., Al-Khulifi, F., Al-Lahou, B., Al-Othman, A., & Davidsson, L. (2015). Total body water measurement using the 2H dilution technique for the assessment of body composition of Kuwaiti children. *Public Health Nutrition*, *18*(02), 259–263. https://doi.org/10.1017/S1368980013003534

Alpert, S. S. (1990). Growth, thermogenesis, and hyperphagia. *American Journal of Clinical Nutrition*, *52*(5), 784–792. https://doi.org/10.1093/ajcn/52.5.784

Alsubheen, S. A., George, A. M., Baker, A., Rohr, L. E., & Basset, F. A. (2016). Accuracy of the vivofit activity tracker. *Journal of Medical Engineering and Technology*, *40*(6), 298–306. https://doi.org/10.1080/03091902.2016.1193238

Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, *547*(7663), 336–339. https://doi.org/10.1038/nature23018

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: the Analysis of Method Comparison Studies †. In *The Statistician* (Vol. 32). http://people.stat.sfu.ca/~raltman/stat300/AltmanBland.pdf

Amati, F., Dubé, J. J., Shay, C., & Goodpaster, B. H. (2008). Separate and combined effects of exercise training and weight loss on exercise efficiency and substrate oxidation. *Journal of Applied Physiology*. https://doi.org/10.1152/japplphysiol.90384.2008

Amatruda, J. M., Statt, M. C., & Welle, S. L. (1993). Total and resting energy expenditure in obese women reduced to ideal body weight. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI116695

Archer, E., Hand, G. A., & Blair, S. N. (2013). Validity of U.S. nutritional surveillance:National Health and Nutrition Examination Survey caloric energy intake data, 1971-2010. *PloS One*, *8*(10), e76632. https://doi.org/10.1371/journal.pone.0076632

Asbeck, I., Mast, M., Bierwag, A., Westenhofer, J., Acheson, K. J., & Muller, M. J. (2002). Severe underreporting of energy intake in normal weight subjects: use of an appropriate standard and relation to restrained eating. *Public Health Nutrition*, *5*(5), 683–690. https://doi.org/10.1079/PHN2002337

Aspvik, N. P., Viken, H., Ingebrigtsen, J. E., Zisko, N., Mehus, I., Wisløff, U., & Stensvold, D. (2018). Do weather changes influence physical activity level among older adults? – The Generation 100 study. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0199463

Assah, F. K., Ekelund, U., Brage, S., Wright, A., Mbanya, J. C., & Wareham, N. J. (2011). Accuracy and validity of a combined heart rate and motion sensor for the measurement of free-living physical activity energy expenditure in adults in Cameroon. *International Journal of Epidemiology*, *40*(1), 112–120. https://doi.org/10.1093/ije/dyq098

Bai, J., Di, C., Xiao, L., Evenson, K. R., LaCroix, A. Z., Crainiceanu, C. M., & Buchner, D. M. (2016). An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics. *PloS One*, *11*(8), e0160644. https://doi.org/10.1371/journal.pone.0160644

Bai, Y., Hibbing, P., Mantis, C., & Welk, G. J. (2018). Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. *Journal of Sports Sciences*, *36*(15), 1734–1741. https://doi.org/10.1080/02640414.2017.1412235

Baracos, V., Caserotti, P., Earthman, C. P., Fields, D., Gallagher, D., Hall, K. D., Heymsfield, S. B., Müller, M. J., Rosen, A. N., Pichard, C., Redman, L. M., Shen, W., Shepherd, J. A., & Thomas, D. (2012). Advances in the Science and Application of Body Composition Measurement. *Journal of Parenteral and Enteral Nutrition*, *36*(1), 96–107. https://doi.org/10.1177/0148607111417448

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479. https://doi.org/10.1037/0003-066X.54.7.462

Bassett, D. R., Rowlands, A., & Trost, S. G. (2012). Calibration and

validation of wearable monitors. *Medicine and Science in Sports and Exercise*, *44*(SUPPL. 1), S32–S38. https://doi.org/10.1249/MSS.0b013e3182399cf7

Bassett, D. R., Wyatt, H. R., Thompson, H., Peters, J. C., & Hill, J. O. (2010). Pedometer-measured physical activity and health behaviors in U.S. adults. *Medicine and Science in Sports and Exercise*, *42*(10), 1819–1825. https://doi.org/10.1249/MSS.0b013e3181dc2e54

Bastian, T., Maire, A., Dugas, J., Ataya, A., Villars, C., Gris, F., Perrin, E., Caritu, Y., Doron, M., Blanc, S., Jallon, P., & Simon, C. (2015). Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: Laboratory-based calibrations are not enough. *Journal of Applied Physiology*, *118*(6), 716–722. https://doi.org/10.1152/japplphysiol.01189.2013

Beaulieu, K., Hopkins, M., Blundell, J., & Finlayson, G. (2017). Impact of physical activity level and dietary fat content on passive overconsumption of energy in non-obese adults. *International Journal of Behavioral Nutrition and Physical Activity*, *14*(1), 14. https://doi.org/10.1186/s12966-017-0473-3

Beaulieu, K., Hopkins, M., Blundell, J., & Finlayson, G. (2018). Homeostatic and non-homeostatic appetite control along the spectrum of physical activity levels: An updated perspective. *Physiology and Behavior*, *192*(December 2017), 23–29. https://doi.org/10.1016/j.physbeh.2017.12.032

Benedetto, S., Caldato, C., Bazzan, E., Greenwood, D. C., Pensabene, V., & Actis, P. (2018). Assessment of the fitbit charge 2 for monitoring heart rate. *PLoS ONE*, *13*(2), e0192691. https://doi.org/10.1371/journal.pone.0192691

Benito, P. J., Neiva, C., González-Quijano, P. S., Cupeiro, R., Morencos, E., & Peinado, A. B. (2012). Validation of the SenseWear armband in circuit resistance training with different loads. *European Journal of Applied Physiology*, *112*(8), 3155–3159. https://doi.org/10.1007/s00421-011-2269-5

Berkemeyer, K., Wijndaele, K., White, T., Cooper, A. J. M., Luben, R., Westgate, K., Griffin, S. J., Khaw, K. T., Wareham, N. J., & Brage, S. (2016). The descriptive epidemiology of accelerometer-measured physical activity in older adults. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1), 1–10. https://doi.org/10.1186/s12966-015-0316-z

Berman, E. S. F., Fortson, S. L., Snaith, S. P., Gupta, M., Baer, D. S., Chery, I., Blanc, S., Melanson, E. L., Thomson, P. J., & Speakman, J. R. (2012). Direct analysis of δ2H and δ18O in natural and enriched human urine using laser-based, off-axis integrated cavity output spectroscopy. *Analytical Chemistry*, *84*(22), 9768–9773. https://doi.org/10.1021/ac3016642

Berman, E. S. F., Swibas, T., Kohrt, W. M., Catenacci, V. A., Creasy, S. A., Melanson, E. L., & Speakman, J. R. (2020). Maximizing precision and accuracy of the doubly labeled water method via optimal sampling

protocol, calculation choices, and incorporation of 17O measurements. *European Journal of Clinical Nutrition*, *74*(3), 454–464. https://doi.org/10.1038/s41430-019-0492-z

Berntsen, S., Hageberg, R., Aandstad, A., Mowinckel, P., Anderssen, S. A., Carlsen, K. H., & Andersen, L. B. (2010). Validity of physical activity monitors in adults participating in free-living activities. *British Journal of Sports Medicine*, *44*(9), 657–664. https://doi.org/10.1136/bjsm.2008.048868

Berntsen, Sveinung, Stafne, S. N., & Maørkved, S. (2011). Physical activity monitor for recording energy expenditure in pregnancy. *Acta Obstetricia et Gynecologica Scandinavica*, *90*(8), 903–907. https://doi.org/10.1111/j.1600-0412.2011.01172.x

Berry, S. E., Valdes, A. M., Drew, D. A., Asnicar, F., Mazidi, M., Wolf, J., Capdevila, J., Hadjigeorgiou, G., Davies, R., Al Khatib, H., Bonnett, C., Ganesh, S., Bakker, E., Hart, D., Mangino, M., Merino, J., Linenberg, I., Wyatt, P., Ordovas, J. M., … Spector, T. D. (2020). Human postprandial responses to food and potential for precision nutrition. *Nature Medicine*, *26*(6), 964–973. https://doi.org/10.1038/s41591-020-0934-0

Bhammar, D. M., Sawyer, B. J., Tucker, W. J., Lee, J.-M., & Gaesser, G. A. (2016). Validity of SenseWear(R) Armband v5.2 and v2.2 for estimating energy expenditure. *Journal of Sports Sciences*, *34*(19), 1830–1838. https://doi.org/10.1080/02640414.2016.1140220

Bhutani, S., Kahn, E., Tasali, E., & Schoeller, D. A. (2017). Composition of two-week change in body weight under unrestricted free-living conditions. *Physiological Reports*, *15*(13). https://doi.org/10.14814/phy2.13336

Black, A. E., & Cole, T. J. (2000). Within- and between-subject variation in energy expenditure measured by the doubly-labelled water technique: Implications for validating reported dietary energy intake. *European Journal of Clinical Nutrition*, *54*(5), 386–394. https://doi.org/10.1038/sj.ejcn.1600970

Black, A. E., Coward, W. A., Cole, T. J., & Prentice, A. M. (1996). Human energy expenditure in affluent societies : An analysis of 574 doubly-labelled water measurements. *European Journal of Clinical Nutrition*, *50*(2), 72–92. http://www.ncbi.nlm.nih.gov/pubmed/8641250

Black, A. E., Prentice, A. M., & Coward, W. A. (1986). Use of food quotients to predict respiratory quotients for the doubly-labelled water method of measuring energy expenditure. *Human Nutrition: Clinical Nutrition*, *40*(5), 381–391.

Blair, C. K., Morey, M. C., Desmond, R. A., Cohen, H. J., Sloane, R., Snyder, D. C., & Demark-Wahnefried, W. (2014). Light-intensity activity attenuates functional decline in older cancer survivors. *Medicine and Science in Sports and Exercise*, *46*(7), 1375–1383. https://doi.org/10.1249/MSS.0000000000000241

Bogardus, C., Lillioja, S., Ravussin, E., Abbott, W., Zawadzki, J. K., Young, A., Knowler, W. C., Jacobowitz, R., & Moll, P. P. (1986). Familial Dependence of the Resting Metabolic Rate. *New England Journal of*

*Medicine*, *315*(2), 96–100. https://doi.org/10.1056/nejm198607103150205

Bonomi, A. G., Goldenberg, S., Papini, G., Kraal, J., Stut, W., Sartor, F., & Kemps, H. (2015). Predicting energy expenditure from photo-plethysmographic measurements of heart rate under beta blocker therapy: Data driven personalization strategies based on mixed models. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2015-Novem*, 7642–7646. https://doi.org/10.1109/EMBC.2015.7320162

Borghese, M. M., Borgundvaag, E., McIsaac, M. A., & Janssen, I. (2019). Imputing accelerometer nonwear time in children influences estimates of sedentary time and its associations with cardiometabolic risk. *International Journal of Behavioral Nutrition and Physical Activity*. https://doi.org/10.1186/s12966-019-0770-0

Boudreaux, B. D., Hebert, E. P., Hollander, D. B., Williams, B. M., Cormier, C. L., Naquin, M. R., Gillan, W. W., Gusew, E. E., & Kraemer, R. R. (2018). Validity of Wearable Activity Monitors during Cycling and Resistance Exercise. *Medicine and Science in Sports and Exercise*, *50*(3), 624–633. https://doi.org/10.1249/MSS.0000000000001471

Bradbury, K. E., Guo, W., Cairns, B. J., Armstrong, M. E. G., & Key, T. J. (2017). Association between physical activity and body fat percentage, with adjustment for BMI: a large cross-sectional analysis of UK Biobank. *BMJ Open*, *7*(3), e011843. https://doi.org/10.1136/bmjopen-2016-011843

Brage, S., Brage, N., Franks, P. W., Ekelund, U., & Wareham, N. J. (2005). Reliability and validity of the combined heart rate and movement sensor actiheart. *European Journal of Clinical Nutrition*, *59*(4), 561–570. https://doi.org/10.1038/sj.ejcn.1602118

Brage, S., Brage, N., Franks, P. W., Ekelund, U., Wong, M. Y., Andersen, L. B., Froberg, K., & Wareham, N. J. (2004). Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. *Journal of Applied Physiology*, *96*(1), 343–351. https://doi.org/10.1152/japplphysiol.00703.2003

Brage, S., Ekelund, U., Brage, N., Hennings, M. A., Froberg, K., Franks, P. W., & Wareham, N. J. (2007). Hierarchy of individual calibration levels for heart rate and accelerometry to measure physical activity. *Journal of Applied Physiology*, *103*(2), 682–692. https://doi.org/10.1152/japplphysiol.00092.2006

Brage, S., Westgate, K., Franks, P. W., Stegle, O., Wright, A., Ekelund, U., & Wareham, N. J. (2015). Estimation of free-living energy expenditure by heart rate and movement sensing: A doubly-labelled water study. *PLoS ONE*, *10*(9), e0137206. https://doi.org/10.1371/journal.pone.0137206

Bray, G. A., & Bouchard, C. (2020). The biology of human overfeeding: A systematic review. *Obesity Reviews*, *21*(9), 1–78. https://doi.org/10.1111/obr.13040

Bray, G. A., Flatt, J. P., Volaufova, J., DeLany, J. P., & Champagne, C. M.

(2008). Corrective responses in human food intake identified from an analysis of 7-d food-intake records. *American Journal of Clinical Nutrition*. https://doi.org/10.3945/ajcn.2008.26289

Brazeau, A. S., Beaudoin, N., Bélisle, V., Messier, V., Karelis, A. D., & Rabasa-Lhoret, R. (2016). Validation and reliability of two activity monitors for energy expenditure assessment. *Journal of Science and Medicine in Sport*, *19*(1), 46–50. https://doi.org/10.1016/j.jsams.2014.11.001

Brazeau, A. S., Karelis, A. D., Mignault, D., Lacroix, M. J., Prudhomme, D., & Rabasa-Lhoret, R. (2011). Accuracy of the SenseWear Armband??? during ergocycling. *International Journal of Sports Medicine*, *32*(10), 761–764. https://doi.org/10.1055/s-0031-1279768

Brazeau, A. S., Suppere, C., Strychar, I., Belisle, V., Demers, S. P., & Rabasa-Lhoret, R. (2014). Accuracy of energy expenditure estimation by activity monitors differs with ethnicity. *International Journal of Sports Medicine*, *35*(10), 847–850. https://doi.org/10.1055/s-0034-1371837

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Browning, R. C., Baker, E. A., Herron, J. A., & Kram, R. (2006). Effects of obesity and sex on the energetic cost and preferred speed of walking. *Journal of Applied Physiology*, *100*(2), 390–398. https://doi.org/10.1152/japplphysiol.00767.2005

Browning, R. C., Reynolds, M. M., Board, W. J., Walters, K. A., & Reiser, R. F. (2013). Obesity does not impair walking economy across a range of speeds and grades. *Journal of Applied Physiology*. https://doi.org/10.1152/japplphysiol.00765.2012

Brožek, J., Grande, F., Anderson, J. T., & Keys, A. (1963). DENSITOMETRIC ANALYSIS OF BODY COMPOSITION: REVISION OF SOME QUANTITATIVE ASSUMPTIONS. *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/j.1749-6632.1963.tb17079.x

Brugniaux, J. V., Niva, A., Pulkkinen, I., Laukkanen, R. M. T., Richalet, J.-P. P., & Pichon, A. P. (2010). Polar activity watch 200: A new device to accurately assess energy expenditure. *British Journal of Sports Medicine*, *44*(4), 245–249. https://doi.org/10.1136/bjsm.2007.045575

Burton, N. W., Khan, A., & Brown, W. J. (2012). How, where and with whom? Physical activity context preferences of three adult groups at risk of inactivity. *British Journal of Sports Medicine*, *46*(16), 1125–1131. https://doi.org/10.1136/bjsports-2011-090554

Büsing, F., Hägele, F. A., Nas, A., Hasler, M., Müller, M. J., & Bosy-Westphal, A. (2019). Impact of energy turnover on the regulation of glucose homeostasis in healthy subjects. *Nutrition and Diabetes*, *9*(1). https://doi.org/10.1038/s41387-019-0089-6

Butland, B., Jebb, S., Kopelman, P., McPherson, K., Thomas, S., Mardell, J., & Parry, V. (2007). Foresight Tackling Obesities: Future Choices – Project report. *Government Office for Science*, 1–161.

https://doi.org/https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/287937/07-1184x-tackling-obesities-future-choices-report.pdf

Byrne, N. M., & Hills, A. P. (2018). How much exercise should be promoted to raise total daily energy expenditure and improve health? *Obesity Reviews*, *19*, 14–23. https://doi.org/10.1111/obr.12788

Byrne, N. M., Hills, A. P., Hunter, G. R., Weinsier, R. L., & Schutz, Y. (2005). Metabolic equivalent: One size does not fit all. *Journal of Applied Physiology*, *99*(3), 1112–1119. https://doi.org/10.1152/japplphysiol.00023.2004

Byrne, N. M., Sainsbury, A., King, N. A., Hills, A. P., & Wood, R. E. (2018). Intermittent energy restriction improves weight loss efficiency in obese men: The MATADOR study. *International Journal of Obesity*, *42*(2), 129–138. https://doi.org/10.1038/ijo.2017.206

Calabro, M. A., Kim, Y., Franke, W. D., Stewart, J. M., & Welk, G. J. (2015). Objective and subjective measurement of energy expenditure in older adults: A doubly labeled water study. *European Journal of Clinical Nutrition*, *69*(7), 850–855. https://doi.org/10.1038/ejcn.2014.241

Calabró, M. A., Lee, J. M., Saint-Maurice, P. F., Yoo, H., & Welk, G. J. (2014). Validity of physical activity monitors for assessing lower intensity activity in adults. *International Journal of Behavioral Nutrition and Physical Activity*, *11*(1), 119. https://doi.org/10.1186/s12966-014-0119-7

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., & Xie, Z. (2018). Deep Learning and Its Applications in Biomedicine. In *Genomics, Proteomics and Bioinformatics* (Vol. 16, Issue 1, pp. 17–32). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2017.07.003

Casiraghi, F., Lertwattanarak, R., Luzi, L., Chavez, A. O., Davalli, A. M., Naegelin, T., Comuzzie, A. G., Frost, P., Musi, N., & Folli, F. (2013). Energy Expenditure Evaluation in Humans and Non-Human Primates by SenseWear Armband. Validation of Energy Expenditure Evaluation by SenseWear Armband by Direct Comparison with Indirect Calorimetry. *PLoS ONE*, *8*(9), e73651. https://doi.org/10.1371/journal.pone.0073651

Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercuse and physical fitness: definitions and distinctions for health-related research. *Public Health Report*.

Catellier, D. J., Hannan, P. J., Murray, D. M., Addy, C. L., Conway, T. L., Yang, S., & Rice, J. C. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Medicine and Science in Sports and Exercise*, *37*(11 SUPPL.). https://doi.org/10.1249/01.mss.0000185651.59486.4e

Catenacci, V. A., Grunwald, G. K., Ingebrigtsen, J. P., Jakicic, J. M., McDermott, M. D., Phelan, S., Wing, R. R., Hill, J. O., & Wyatt, H. R. (2011). Physical activity patterns using accelerometry in the national weight control registry. *Obesity*, *19*(6), 1163–1170. https://doi.org/10.1038/oby.2010.264

Ceesay, S. M., Prentice, A. M., Day, K. C., Murgatroyd, P. R., Goldberg, G.

R., Scott, W., & Spurr, G. B. (1989). The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry. *British Journal of Nutrition*, *61*(2), 175–186. https://doi.org/10.1079/bjn19890107

Celis-Morales, C. A., Perez-Bravo, F., Ibañez, L., Salas, C., Bailey, M. E. S., & Gill, J. M. R. (2012). Objective vs. self-reported physical activity and sedentary time: Effects of measurement method on relationships with risk biomarkers. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0036345

Chan, C. B., Ryan, D. A. J., & Tudor-Locke, C. (2006). Relationship between objective measures of physical activity and weather: A longitudinal study. *International Journal of Behavioral Nutrition and Physical Activity*, *3*. https://doi.org/10.1186/1479-5868-3-21

Chearskul, S., Delbridge, E., Shulkes, A., Proietto, J., & Kriketos, A. (2008). Effect of weight loss and ketosis on postprandial cholecystokinin and free fatty acid concentrations. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/87.5.1238

Chen, C., Jerome, G. J., Laferriere, D., Young, D. R., & Vollmer, W. M. (2009). Procedures used to standardize data collected by RT3 triaxial accelerometers in a large-scale weight-loss trial. *Journal of Physical Activity & Health*, *6*(3), 354–359. http://www.ncbi.nlm.nih.gov/pubmed/19564665

Chen, K. Y., Acra, S. A., Donahue, C. L., Sun, M., & Buchowski, M. S. (2004). Efficiency of walking and stepping: Relationship to body fatness. *Obesity Research*, *12*(6), 982–989. https://doi.org/10.1038/oby.2004.120

Chen, K. Y., Acra, S. A., Majchrzak, K., Donahue, C. L., Baker, L., Clemens, L., Sun, M., & Buchowski, M. S. (2003). Predicting Energy Expenditure of Physical Activity Using Hip- and Wrist-Worn Accelerometers. *Diabetes Technology and Therapeutics*, *5*(6), 1023–1033. https://doi.org/10.1089/152091503322641088

Chen, K. Y., & Bassett, D. R. (2005). The technology of accelerometry-based activity monitors: Current and future. *Medicine and Science in Sports and Exercise*, *37*(11 SUPPL.), S490-500. https://doi.org/10.1249/01.mss.0000185571.49104.82

Choi, L., Liu, Z., Matthews, C. E., & Buchowski, M. S. (2011). Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and Science in Sports and Exercise*, *43*(2), 357–364. https://doi.org/10.1249/MSS.0b013e3181ed61a3

Chollet, F. (2015). *GitHub - keras-team/keras: Deep Learning for humans*. https://github.com/keras-team/

Chow, C. C., & Hall, K. D. (2008). The dynamics of human body weight change. *PLoS Computational Biology*, *4*(3), e1000045. https://doi.org/10.1371/journal.pcbi.1000045

Chow, C. C., & Hall, K. D. (2014). Short and long-term energy intake patterns and their implications for human body weight regulation.

*Physiology and Behavior*, *134*(C), 60–65. https://doi.org/10.1016/j.physbeh.2014.02.044

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry. In *Medicine and Science in Sports and Exercise* (Vol. 49, Issue 9). Lippincott Williams and Wilkins. https://doi.org/10.1249/MSS.0000000000001291

Chowdhury, E. A., Western, M. J., Nightingale, T. E., Peacock, O. J., & Thompson, D. (2017). Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors. *PLoS ONE*, *12*(2), e0171720. https://doi.org/10.1371/journal.pone.0171720

Chumlea, W. C. (2006). Body Composition Assessment of Obesity. In *Overweight and the metabolic syndrome: from bench to bedside.* (pp. 23–35.). Springer US. https://doi.org/10.1007/978-0-387-32164-6_2

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press. https://www.sciencedirect.com/science/book/9780121790608

Colbert, L. H., Matthews, C. E., Havighurst, T. C., Kim, K., & Schoeller, D. A. (2011). Comparative validity of physical activity measures in older adults. *Medicine and Science in Sports and Exercise*, *43*(5), 867–876. https://doi.org/10.1249/MSS.0b013e3181fc7162

Cole, J. B., Florez, J. C., & Hirschhorn, J. N. (2020). Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic associations. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15193-0

Cole, T. J., & Coward, W. A. (1992). Precision and accuracy of doubly labeled water energy expenditure by multipoint and two-point methods. *The American Journal of Physiology*, *263*(5 Pt 1), E965-73. https://doi.org/10.1152/ajpendo.1992.263.5.E965

Cooper, G. M. (2018). *The Cell: A Molecular Approach* (8th ed.). Oxford University Press. https://learninglink.oup.com/access/cooper8e

Correa, J. B., Apolzan, J. W., Shepard, D. N., Heil, D. P., Rood, J. C., & Martin, C. K. (2016). Evaluation of the ability of three physical activity monitors to predict weight change and estimate energy expenditure. *Applied Physiology, Nutrition, and Metabolism = Physiologie Appliquee, Nutrition et Metabolisme*, *41*(7), 758–766. https://doi.org/10.1139/apnm-2015-0461

Crouter, S. E., Churilla, J. R., & Bassett, D. R. (2006). Estimating energy expenditure using accelerometers. *European Journal of Applied Physiology*, *98*(6), 601–612. https://doi.org/10.1007/s00421-006-0307-5

Crouter, S. E., Churilla, J. R., & Bassett, D. R. (2008). Accuracy of the Actiheart for the assessment of energy expenditure in adults. *European Journal of Clinical Nutrition*, *62*(6), 704–711. https://doi.org/10.1038/sj.ejcn.1602766

Crouter, S. E., Kuffel, E., Haas, J. G. D., Frongillo, E. A., Bassett, D. R. J.,

Frongllo, E. A., & Bassett, D. R. J. (2010). Refined two-regression model for the ActiGraph accelerometer. *Medicine & Science in Sports & Exercise*, *42*(5), 1029–1037. https://doi.org/10.1249/MSS.0b013e3181c37458

Cummings, D. E., Weigle, D. S., Frayo, R. S., Breen, P. A., Ma, M. K., Dellinger, E. P., & Purnell, J. Q. (2002). Plasma Ghrelin Levels after Diet-Induced Weight Loss or Gastric Bypass Surgery. *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa012908

da Rocha, E. E. M., Alves, V. G. F., & da Fonseca, R. B. V. (2006). Indirect calorimetry: methodology, instruments and clinical application. *Current Opinion in Clinical Nutrition and Metabolic Care*, *9*(3), 247–256. https://doi.org/10.1097/01.mco.0000222107.15548.f5

Das, S. K., Roberts, S. B., McCrory, M. A., George Hsu, L. K., Shikora, S. A., Kehayias, J. J., Dallal, G. E., & Saltzman, E. (2003). Long-term changes in energy expenditure and body composition after massive weight loss induced by gastric bypass surgery 1-4. *American Journal of Clinical Nutrition*, *78*(1), 22–30. https://doi.org/10.1093/ajcn/78.1.22

de Almeida Mendes, M., da Silva, I., Ramires, V., Reichert, F., Martins, R., Ferreira, R., & Tomasi, E. (2018). Metabolic equivalent of task (METs) thresholds as an indicator of physical activity intensity. *PLoS ONE*, *13*(7). https://doi.org/10.1371/journal.pone.0200701

de Arriba-Pérez, F., Caeiro-Rodríguez, M., & Santos-Gago, J. M. (2016). Collection and processing of data from wrist wearable devices in heterogeneous and multiple-user scenarios. *Sensors (Switzerland)*, *16*(9). https://doi.org/10.3390/s16091538

de Castro, J. M. (2010). The control of food intake of free-living humans: Putting the pieces back together. *Physiology and Behavior*. https://doi.org/10.1016/j.physbeh.2010.04.028

de Castro, J. M., & Plunkett, S. (2002). A general model of intake regulation. In *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/S0149-7634(02)00018-0

de Jonge, L, & Bray, G. A. (1997). The thermic effect of food and obesity: a critical review. *Obesity Research*, *5*(6), 622–631. http://www.ncbi.nlm.nih.gov/pubmed/9449148

de Jonge, Lilian, DeLany, J. P., Nguyen, T., Howard, J., Hadley, E. C., Redman, L. M., & Ravussin, E. (2007). Validation study of energy expenditure and intake during calorie restriction using doubly labeled water and changes in body composition. *The American Journal of Clinical Nutrition*, *85*(1), 73–79. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2662402&tool=pmcentrez&rendertype=abstract

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M., Altman, D. G., International Stroke Trial Collaborative Group, & European Carotid Surgery Trial Collaborative Group. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment (Winchester, England)*, *7*(27), iii–x, 1–173. http://www.ncbi.nlm.nih.gov/pubmed/14499048

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., Fisher, C. B., Harness, D., Viswanath, S., Heymsfield, S. B., Dungan, I., & Thomas, D. M. (2018). A review of machine learning in obesity. In *Obesity Reviews* (Vol. 19, Issue 5, pp. 668–685). Blackwell Publishing Ltd. https://doi.org/10.1111/obr.12667

Degroote, L., Desmet, A., De Bourdeaudhuij, I., Van Dyck, D., & Crombez, G. (2020). Content validity and methodological considerations in ecological momentary assessment studies on physical activity and sedentary behaviour: A systematic review. In *International Journal of Behavioral Nutrition and Physical Activity* (Vol. 17, Issue 1). https://doi.org/10.1186/s12966-020-00932-9

Delany, J. P. (2012). Measurement of energy expenditure. In *Pediatric Blood and Cancer* (Vol. 58, Issue 1, pp. 129–134). https://doi.org/10.1002/pbc.23369

Dempster, P., & Aitkens, S. (1995). A new air displacement method for the determination of human body composition. *Medicine and Science in Sports and Exercise*, *27*(12), 1692–1697. http://www.ncbi.nlm.nih.gov/pubmed/8614327

Dhurandhar, N. V., Schoeller, D., Brown, A. W., Heymsfield, S. B., Thomas, D., Sørensen, T. I. A., Speakman, J. R., Jeansonne, M., & Allison, D. B. (2015). Energy balance measurement: When something is not better than nothing. *International Journal of Obesity*, *39*(7), 1109–1113. https://doi.org/10.1038/ijo.2014.199

Diaz, K. M., Krupka, D. J., Chang, M. J., Peacock, J., Ma, Y., Goldsmith, J., Schwartz, J. E., & Davidson, K. W. (2016). Fitbit®: an Accurate and Reliable Device for Wireless Physical Activity Tracking. *International Journal of Cardiology*, *185*, 138–140. https://doi.org/10.1016/j.ijcard.2015.03.038.FITBIT

Diaz, K. M., Krupka, D. J., Chang, M. J., Shaffer, J. A., Ma, Y., Goldsmith, J., Schwartz, J. E., & Davidson, K. W. (2016). Validation of the Fitbit One® for physical activity measurement at an upper torso attachment site. *BMC Research Notes*, *9*(1), 213. https://doi.org/10.1186/s13104-016-2020-8

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S., & Wareham, N. J. (2017). Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLoS ONE*, *12*(2), e0169649. https://doi.org/10.1371/journal.pone.0169649

Doherty, A. R., Kelly, P., Kerr, J., Marshall, S., Oliver, M., Badland, H., Hamilton, A., & Foster, C. (2013). Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 22. https://doi.org/10.1186/1479-5868-10-22

Dohrn, I. M., Sjöström, M., Kwak, L., Oja, P., & Hagströmer, M. (2018). Accelerometer-measured sedentary time and physical activity—A 15 year follow-up of mortality in a Swedish population-based cohort.

*Journal of Science and Medicine in Sport*. https://doi.org/10.1016/j.jsams.2017.10.035

Dondzila, C., & Garner, D. (2016). Comparative accuracy of fitness tracking modalities in quantifying energy expenditure. *Journal of Medical Engineering and Technology*, *40*(6), 325–329. https://doi.org/10.1080/03091902.2016.1197978

Dooley, E. E., Golaszewski, N. M., & Bartholomew, J. B. (2017). Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. *JMIR MHealth and UHealth*, *5*(3), e34. https://doi.org/10.2196/mhealth.7043

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, *52*(6), 377–384. https://doi.org/10.1136/JECH.52.6.377

Drenowatz, C., & Eisenmann, J. C. (2011). Validation of the SenseWear Armband at high intensity exercise. *European Journal of Applied Physiology*, *111*(5), 883–887. https://doi.org/10.1007/s00421-010-1695-0

Dulloo, A. G. (2010). Energy Balance and Body Weight Homeostasis. In *Clinical Obesity in Adults and Children* (pp. 65–81). https://doi.org/10.1002/9781444307627.ch6

Dulloo, A. G., Jacquet, J., Montani, J. P., & Schutz, Y. (2012). Adaptive thermogenesis in human body weight regulation: More of a concept than a measurable entity? *Obesity Reviews*, *13*(SUPPL.2), 105–121. https://doi.org/10.1111/j.1467-789X.2012.01041.x

Duren, D. L., Sherwood, R. J., Czerwinski, S. A., Lee, M., Choh, A. C., Siervogel, R. M., & Cameron Chumlea, W. (2008). Body composition methods: comparisons and interpretation. *Journal of Diabetes Science and Technology*, *2*(6), 1139–1146. https://doi.org/10.1177/193229680800200623

Durnin, J. V. G. A. (1961). Basic physiological factors affecting calorie balance. *Proceedings of the Nutrition Society*, *20*(1), 52–58. https://doi.org/10.1079/pns19610013

Durnin, J. V. G. A. (1991). Practical Estimates of Energy Requirements. *Journal of Nutrition*, *121*, 1907–1913.

Edholm, O. G., Adam, J. M., Healy, M. J. R., Wolff, H. S., Goldsmith, R., & Best, T. W. (1970). Food intake and energy expenditure of army recruits. *British Journal of Nutrition*, *24*(4), 1091–1107. https://doi.org/10.1079/bjn19700112

Edholm, O. G., Fletcher, J. G., Widdowson, E. M., & McCance, R. A. (1955). The Energy Expenditure and Food Intake of Individual Men. *British Journal of Nutrition*, *9*(3), 286–300. https://doi.org/10.1079/bjn19550040

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Ekelund, U., Tarp, J., Steene-Johannessen, J., Hansen, B. H., Jefferis, B., Fagerland, M. W., Whincup, P., Diaz, K. M., Hooker, S. P., Chernofsky, A., Larson, M. G., Spartano, N., Vasan, R. S., Dohrn, I. M., Hagströmer, M., Edwardson, C., Yates, T., Shiroma, E., Anderssen, S. A., & Lee, I. M. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: Systematic review and harmonised meta-analysis. *The BMJ*. https://doi.org/10.1136/bmj.l4570

Ekelund, U., Yngve, A., Westerterp, K., & Sjöström, M. (2002). Energy expenditure assessed by heart rate and doubly labeled water in young athletes. *Medicine and Science in Sports and Exercise*, *34*(8), 1360–1366. http://www.ncbi.nlm.nih.gov/pubmed/12165693

Elia, M. (1991). Energy equivalents of CO2 and their importance in assessing energy expenditure when using tracer techniques. *American Journal of Physiology - Endocrinology and Metabolism*, *260*(1 23-1). https://doi.org/10.1152/ajpendo.1991.260.1.e75

Elia, M., & Cummings, J. H. (2007). Physiological aspects of energy metabolism and gastrointestinal effects of carbohydrates. *European Journal of Clinical Nutrition*, *61*, S40–S74. https://doi.org/10.1038/sj.ejcn.1602938

Elia, M., & Livesey, G. (1992). Energy Expenditure and Fuel Selection in Biological Systems: The Theory and Practice of Calculations Based on Indirect Calorimetry and Tracer Methods. *World Review of Nutrition and Dietetics*, *70*(May), 68–131. https://doi.org/10.1159/000421672

Elia, M., Stratton, R., & Stubbs, J. (2003). Techniques for the study of energy balance in man. *Proceedings of the Nutrition Society*. https://doi.org/10.1079/PNS2003255

Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, D., & Marshall, S. (2014). A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological Measurement*, *35*(11), 2191–2203. https://doi.org/10.1088/0967-3334/35/11/2191

Ellis, K., Kerr, J., Godbole, S., Staudenmayer, J., & Lanckriet, G. (2016). Hip and wrist accelerometer algorithms for free-living behavior classification. *Medicine and Science in Sports and Exercise*, *48*(5), 933–940. https://doi.org/10.1249/MSS.0000000000000840

Ells, L. J., Demaio, A., & Farpour-Lambert, N. (2018). Diet, genes, and obesity: Genetic predisposition to obesity is no barrier to successful weight management. *BMJ (Online)*, *360*, k7. https://doi.org/10.1136/bmj.k7

Erdogan, A., Cetin, C., Karatosun, H., & Baydar, M. L. (2010). Accuracy of the Polar S810i(TM) Heart Rate Monitor and the Sensewear Pro Armband(TM) to Estimate Energy Expenditure of Indoor Rowing Exercise in Overweight and Obese Individuals. *Journal of Sports Science & Medicine*, *9*(3), 508–516.

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers.

*International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 159. https://doi.org/10.1186/s12966-015-0314-1

FAO/WHO/UNU. (2004). Human energy requirements. Report of a Joint FAO/WHO/UNU Expert Consultation. *AO Food and Nutrition Technical Report Series*.

Farooqi, N., Slinde, F., Håglin, L., & Sandström, T. (2013). Validation of SenseWear Armband and ActiHeart monitors for assessments of daily energy expenditure in free-living women with chronic obstructive pulmonary disease. *Physiological Reports*, *1*(6), n/a-n/a. https://doi.org/10.1002/phy2.150

Farrahi, V., Niemelä, M., Kangas, M., Korpelainen, R., & Jämsä, T. (2019). Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches. *Gait and Posture*, *68*(December 2018), 285–299. https://doi.org/10.1016/j.gaitpost.2018.12.003

Farrahi, V., Niemela, M., Tjurin, P., Kangas, M., Korpelainen, R., & Jamsa, T. (2020). Evaluating and Enhancing the Generalization Performance of Machine Learning Models for Physical Activity Intensity Prediction from Raw Acceleration Data. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 27–38. https://doi.org/10.1109/JBHI.2019.2917565

Feehan, L. M., Geldman, J., Sayre, E. C., Park, C., Ezzat, A. M., Young Yoo, J., Hamilton, C. B., & Li, L. C. (2018). Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. *JMIR MHealth and UHealth*, *6*(8), e10527. https://doi.org/10.2196/10527

Ferrannini, E. (1988). The theoretical bases of indirect calorimetry: A review. *Metabolism*, *37*(3), 287–301. https://doi.org/10.1016/0026-0495(88)90110-2

Fields, D. A., Goran, M. I., & McCrory, M. A. (2002). Body-composition assessment via air-displacement plethysmography in adults and children: A review. *American Journal of Clinical Nutrition*, *75*(3), 453–467. https://doi.org/10.1093/ajcn/75.3.453

Flack, K. D., Siders, W. A., Johnson, L. A., & Roemmich, J. N. (2016). Cross-Validation of Resting Metabolic Rate Prediction Equations. *Journal of the Academy of Nutrition and Dietetics*, *116*(9), 1413–1422. https://doi.org/10.1016/j.jand.2016.03.018

Flegal, K. M., Kit, B. K., Orpana, H., & Graubard, B. I. (2013). Association of all-cause mortality with overweight and obesity using standard body mass index categories a systematic review and meta-analysis. *JAMA - Journal of the American Medical Association*, *309*(1), 71–82. https://doi.org/10.1001/jama.2012.113905

Fletcher, G. F., Balady, G. J., Amsterdam, E. A., Chaitman, B., Eckel, R., Fleg, J., Froelicher, V. F., Leon, A. S., Piña, I. L., Rodney, R., Simons-Morton, D. A., Williams, M. A., & Bazzarre, T. (2001). Exercise Standards for Testing and Training. *Circulation*, *104*(14), 1694–1740. https://doi.org/10.1161/hc3901.095960

Forbes, G. B. (1987). Lean body mass-body fat interrelationships in humans.

*Nutrition Reviews*, *45*(8), 225–231. http://www.ncbi.nlm.nih.gov/pubmed/3306482

Forbes, G. B. (2000). Body fat content influences the body composition response to nutrition and exercise. *Annals of the New York Academy of Sciences*, *904*, 359–365. https://doi.org/10.1111/j.1749-6632.2000.tb06482.x

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2020). Sharpness-Aware Minimization for Efficiently Improving Generalization. *ArXiv*. http://arxiv.org/abs/2010.01412

Foright, R. M., Presby, D. M., Sherk, V. D., Kahn, D., Checkley, L. A., Giles, E. D., Bergouignan, A., Higgins, J. A., Jackman, M. R., Hill, J. O., & MacLean, P. S. (2018). Is regular exercise an effective strategy for weight loss maintenance? *Physiology and Behavior*, *188*(January), 86–93. https://doi.org/10.1016/j.physbeh.2018.01.025

Foss, Ø., & Hallén, J. (2005). Validity and stability of a computerized metabolic system with mixing chamber. *International Journal of Sports Medicine*, *26*(7), 569–575. https://doi.org/10.1055/s-2004-821317

Fothergill, E., Guo, J., Howard, L., Kerns, J. C., Knuth, N. D., Brychta, R., Chen, K. Y., Skarulis, M. C., Walter, M., Walter, P. J., & Hall, K. D. (2016). Persistent metabolic adaptation 6 years after “The Biggest Loser” competition. *Obesity*, *24*(8), 1612–1619. https://doi.org/10.1002/oby.21538

Frankenfield, D., Roth-Yousey, L., & Compher, C. (2005). Comparison of Predictive Equations for Resting Metabolic Rate in Healthy Nonobese and Obese Adults: A Systematic Review. *Journal of the American Dietetic Association*, *105*(5), 775–789. https://doi.org/10.1016/j.jada.2005.02.005

Frayn, K. N., & Evans, R. (2019). *Human Metabolism: A Regulatory Perspective* (4th ed.). Wiley.

Freedman, L. S., Commins, J. M., Moler, J. E., Arab, L., Baer, D. J., Kipnis, V., Midthune, D., Moshfegh, A. J., Neuhouser, M. L., Prentice, R. L., Schatzkin, A., Spiegelman, D., Subar, A. F., Tinker, L. F., & Willett, W. (2014). *Pooled results from 5 validation studies of dietary self-report instruments using recovery biomarkers for energy and protein intake*. American Journal of Epidemiology. https://doi.org/10.1093/aje/kwu116

Freedson, P. S., Melanson, E., & Sirard, J. (1998). Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine and Science in Sports and Exercise*, *30*(5), 777–781.

Fruin, M. L., & Rankin, J. W. (2004). Validity of a multi-sensor Armband in estimating rest and exercise energy expenditure. *Medicine and Science in Sports and Exercise*, *36*(6), 1063–1069. https://doi.org/10.1249/01.MSS.0000128144.91337.38

Furlanetto, K. C., Bisca, G. W., Oldemberg, N., Sant’Anna, T. J., Morakami, F. K., Camillo, C. A., Cavalheri, V., Hernandes, N. A., Probst, V. S., Ramos, E. M., Brunetto, A. F., & Pitta, F. (2010). Step Counting and Energy Expenditure Estimation in Patients With Chronic Obstructive

Pulmonary Disease and Healthy Elderly: Accuracy of 2 Motion Sensors. *Archives of Physical Medicine and Rehabilitation*, *91*(2), 261–267. https://doi.org/10.1016/j.apmr.2009.10.024

Galgani, J., & Ravussin, E. (2008). Energy metabolism, fuel selection and body weight regulation. *International Journal of Obesity*, *32*(SUPPL. 7). https://doi.org/10.1038/ijo.2008.246

Garden, L., Clark, H., Whybrow, S., & Stubbs, R. J. (2018). Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific? *European Journal of Clinical Nutrition*, *72*(7), 1026–1034. https://doi.org/10.1038/s41430-018-0199-6

Garland, T., Schutz, H., Chappell, M. A., Keeney, B. K., Meek, T. H., Copes, L. E., Acosta, W., Drenowatz, C., Maciel, R. C., Van Dijk, G., Kotz, C. M., & Eisenmann, J. C. (2011). The biological control of voluntary exercise, spontaneous physical activity and daily energy expenditure in relation to obesity: Human and rodent perspectives. *Journal of Experimental Biology*, *214*(2), 206–229. https://doi.org/10.1242/jeb.048397

Gastin, P. B., Cayzer, C., Dwyer, D., & Robertson, S. (2018). Validity of the ActiGraph GT3X+ and BodyMedia SenseWear Armband to estimate energy expenditure during physical activity and sport. *Journal of Science and Medicine in Sport*, *21*(3), 291–295. https://doi.org/10.1016/j.jsams.2017.07.022

Geldszus, R., Mayr, B., Horn, R., Geisthövel, F., Von Zur Mühlen, A., & Brabant, G. (1996). Serum leptin and weight reduction in female obesity. *European Journal of Endocrinology*. https://doi.org/10.1530/eje.0.1350659

Genton, L., Hans, D., Kyle, U. G., & Pichard, C. (2002). Dual-energy X-ray absorptiometry and body composition: differences between devices and comparison with reference methods. *Nutrition* , *18*(1), 66–70. https://doi.org/10.1016/S0899-9007(01)00700-6

Géron, A. (2019). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media* (2nd ed.).

Gibbs, B. B., & Davis, K. K. (2018). *In Pursuit of the " Something " that Is Better than Nothing for Measuring Energy Intake*. *March*, 2017–2018. https://doi.org/10.1093/jn/nxy006

Gillinov, S., Etiwy, M., Wang, R., Blackburn, G., Phelan, D., Gillinov, A. M., Houghtaling, P., Javadikasgari, H., & Desai, M. Y. (2017). Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine and Science in Sports and Exercise*, *49*(8), 1697–1703. https://doi.org/10.1249/MSS.0000000000001284

Gilmore, L. A., Ravussin, E., Bray, G. A., Han, H., & Redman, L. M. (2014). An objective estimate of energy intake during weight gain using the intake-balance method. *American Journal of Clinical Nutrition*, *100*(3), 806–812. https://doi.org/10.3945/ajcn.114.087122

Göbel, B., Sanghvi, A., & Hall, K. D. (2014). Quantifying energy intake

changes during obesity pharmacotherapy. *Obesity*, *22*(10), 2105–2108. https://doi.org/10.1002/oby.20813

Goldberg, G., Black, A., Jebb, S., Colte, T., Murgatroyd, P., Coward, W., & Prentice, A. (1991). Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording. *Euro J Clin Nutr*, *45*(12), 569–581. https://doi.org/10.1377/hlthaff.2011.1088

Goldberg, G. R., Prentice, A. M., Davies, H. L., & Murgatroyd, P. R. (1988). Overnight and basal metabolic rates in men and women. *European Journal of Clinical Nutrition*, *42*(2), 137–144. http://www.ncbi.nlm.nih.gov/pubmed/3378547

Gonzalez, J. T., Veasey, R. C., Rumbold, P. L. S., & Stevenson, E. J. (2013). Breakfast and exercise contingently affect postprandial metabolism and energy balance in physically active males. *British Journal of Nutrition*, *110*(4), 721–732. https://doi.org/10.1017/S0007114512005582

Greenway, F. L. (2015). Physiological adaptations to weight loss and factors favouring weight regain. *International Journal of Obesity*, *39*(8), 1188–1196. https://doi.org/10.1038/ijo.2015.59

Gualtieri, L., Rosenbluth, S., & Phillips, J. (2016). Can a Free Wearable Activity Tracker Change Behavior? The Impact of Trackers on Adults in a Physician-Led Wellness Group. *JMIR Research Protocols*, *5*(4), e237. https://doi.org/10.2196/resprot.6534

Guo, J., Robinson, J. L., Gardner, C. D., & Hall, K. D. (2019). Objective versus Self-Reported Energy Intake Changes During Low-Carbohydrate and Low-Fat Diets. *Obesity*, *27*(3), 420–426. https://doi.org/10.1002/oby.22389

Gupta, R. Das, Ramachandran, R., Venkatesan, P., Anoop, S., Joseph, M., & Thomas, N. (2017). Indirect Calorimetry: From Bench to Bedside. *Indian Journal of Endocrinology and Metabolism*, *21*(4), 594–599. https://doi.org/10.4103/ijem.IJEM_484_16

Hägele, F. A., Büsing, F., Nas, A., Hasler, M., Müller, M. J., Blundell, J. E., & Bosy-Westphal, A. (2019). Appetite Control Is Improved by Acute Increases in Energy Turnover at Different Levels of Energy Balance. *Journal of Clinical Endocrinology and Metabolism*. https://doi.org/10.1210/jc.2019-01164

Hall, K. D. (2007). Body fat and fat-free mass inter-relationships: Forbes's theory revisited. *British Journal of Nutrition*, *97*(06), 1059. https://doi.org/10.1017/S0007114507691946

Hall, K. D. (2008). What is the required energy deficit per unit weight loss? *International Journal of Obesity*, *32*(3), 573–576. https://doi.org/10.1038/sj.ijo.0803720

Hall, K. D. (2014). Estimating human energy intake using mathematical models. *American Journal of Clinical Nutrition*, *100*(3), 744–745. https://doi.org/10.3945/ajcn.114.094441

Hall, K. D., & Chow, C. C. (2011). Estimating changes in free-living energy

intake and its confidence interval. *American Journal of Clinical Nutrition*, *94*(1), 66–74. https://doi.org/10.3945/ajcn.111.014399

Hall, K. D., Guo, J., Chen, K. Y., Leibel, R. L., Reitman, M. L., Rosenbaum, M., Smith, S. R., & Ravussin, E. (2019). Methodologic considerations for measuring energy expenditure differences between diets varying in carbohydrate using the doubly labeled water method. *American Journal of Clinical Nutrition*, *109*(5), 1328–1334. https://doi.org/10.1093/ajcn/nqy390

Hall, K. D., Heymsfield, S. B., Kemnitz, J. W., Klein, S., Schoeller, D. A., & Speakman, J. R. (2012). Energy balance and its components: Implications for body weight regulation. *American Journal of Clinical Nutrition*, *95*(4), 989–994. https://doi.org/10.3945/ajcn.112.036350

Hall, K. D., & Jordan, P. N. (2008). Modeling weight-loss maintenance to help prevent body weight regain. *The American Journal of Clinical Nutrition*, *88*(6), 1495–1503. https://doi.org/10.3945/ajcn.2008.26333

Hamner, B., Frasco, M., & Ledell, E. (2018). *Package "Metrics" Title Evaluation Metrics for Machine Learning*. https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

Hand, G. A., & Blair, S. N. (2014). Energy flux and its role in obesity and metabolic disease. *European Endocrinology*, *10*(2), 131–135. https://doi.org/10.17925/EE.2014.10.02.131

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Elements of Statistical Learning 2nd ed. In *Springer Series in Statistics*.

Haugen, H. A., Chan, L.-N., & Li, F. (2007). Indirect Calorimetry: A Practical Guide for Clinicians. *Nutrition in Clinical Practice*, *22*(4), 377–388. https://doi.org/10.1177/0115426507022004377

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107. https://doi.org/10.2307/1164588

Heiermann, S., Khalaj Hedayati, K., Müller, M. J., & Dittmar, M. (2011). Accuracy of a portable multisensor body monitor for predicting resting energy expenditure in older people: A comparison with indirect calorimetry. *Gerontology*, *57*(5), 473–479. https://doi.org/10.1159/000322109

Helmerhorst, H. J. F., Brage, S., Warren, J., Besson, H., & Ekelund, U. (2012). A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 103. https://doi.org/10.1186/1479-5868-9-103

Hendelman, D., Miller, K., Baggett, C., Debold, E., & Freedson, P. (2000). Validity of accelerometry for the assessment of moderate intensity physical activity in the field. *Medicine and Science in Sports and Exercise*, *32*(9 Suppl), S442-9. http://www.ncbi.nlm.nih.gov/pubmed/10993413

Herrmann, S. D., Barreira, T. V., Kang, M., & Ainsworth, B. E. (2013). How many hours are enough? Accelerometer wear time may provide bias in

daily activity estimates. In *Journal of Physical Activity and Health* (Vol. 10, Issue 5). https://doi.org/10.1123/jpah.10.5.742

Heymsfield, S. B., Gonzalez, M. C. C., Shen, W., Redman, L., & Thomas, D. (2014). Weight loss composition is one-fourth fat-free mass: A critical review and critique of this widely cited rule. *Obesity Reviews*, *15*(4), 310–321. https://doi.org/10.1111/obr.12143

Heymsfield, S. B., Peterson, C. M., Thomas, D. M., Hirezi, M., Zhang, B., Smith, S., Bray, G., & Redman, L. (2017). Establishing energy requirements for body weight maintenance: Validation of an intake-balance method NCT01672632 NCT. *BMC Research Notes*, *10*(1), 1–8. https://doi.org/10.1186/s13104-017-2546-4

Heymsfield, S. B., Thomas, D., Bosy-Westphal, A., Shen, W., Peterson, C. M., & Müller, M. J. (2012). Evolving concepts on adjusting human resting energy expenditure measurements for body size. *Obesity Reviews*, *13*(11), 1001–1014. https://doi.org/10.1111/j.1467-789X.2012.01019.x

Heymsfield, S. B., Thomas, D., Nguyen, A. M., Peng, J. Z., Martin, C., Shen, W., Strauss, B., Bosy-Westphal, A., & Muller, M. J. (2011). Voluntary weight loss: Systematic review of early phase body composition changes. *Obesity Reviews*, *12*(5), 348–361. https://doi.org/10.1111/j.1467-789X.2010.00767.x

Heymsfield, S B. (1997). Human body composition: advances in models and methods. *Annual Review of Nutrition*, *2*(1), 1–4. http://www.annualreviews.org/doi/abs/10.1146/annurev.nutr.17.1.527

Heymsfield, Steven B., & Wadden, T. A. (2017). Mechanisms, Pathophysiology, and Management of Obesity. *New England Journal of Medicine*, *376*(3), 254–266. https://doi.org/10.1056/nejmra1514009

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Hill, J. O., Peters, J. C., & Wyatt, H. R. (2009). Using the Energy Gap to Address Obesity: A Commentary. *Journal of the American Dietetic Association*. https://doi.org/10.1016/j.jada.2009.08.007

Hill, J. O., Wyatt, H. R., & Peters, J. C. (2012). Energy balance and obesity. *Circulation*, *126*(1), 126–132. https://doi.org/10.1161/CIRCULATIONAHA.111.087213

Hills, A. P., Mokhtar, N., & Byrne, N. M. (2014). Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Frontiers in Nutrition*, *1*(June), 1–16.

https://doi.org/10.3389/fnut.2014.00005

Hirsch, J., & Knittle, J. L. (1970). Cellularity of obese and nonobese human adipose tissue. *Federation Proceedings*, *29*(4), 1516–1521. http://www.ncbi.nlm.nih.gov/pubmed/5459900

Hjorth, M. F., Ritz, C., Blaak, E. E., Saris, W. H. M., Langin, D., Poulsen, S. K., Larsen, T. M., Sørensen, T. I. A., Zohar, Y., & Astrup, A. (2017). Pretreatment fasting plasma glucose and insulin modify dietary weight loss success: Results from 3 randomized clinical trials. *American Journal of Clinical Nutrition*, *106*(2), 499–505. https://doi.org/10.3945/ajcn.117.155200

Höchsmann, C., Dorling, J. L., Apolzan, J. W., Johannsen, N. M., Hsia, D. S., & Martin, C. K. (2020). Baseline Habitual Physical Activity Predicts Weight Loss, Weight Compensation, and Energy Intake During Aerobic Exercise. *Obesity*, *28*(5), 882–892. https://doi.org/10.1002/oby.22766

Holt, G. M., Owen, L. J., Till, S., Cheng, Y., Grant, V. A., Harden, C. J., & Corfe, B. M. (2017). Systematic literature review shows that appetite rating does not predict energy intake. *Critical Reviews in Food Science and Nutrition*, *57*(16), 3577–3582. https://doi.org/10.1080/10408398.2016.1246414

Holt, S. H. A., Brand Miller, J. C., Petocz, P., & Farmakalidis, E. (1995). A satiety index of common foods. *European Journal of Clinical Nutrition*, *49*(9), 675–690.

Hopkins, M., & Blundell, J. E. (2016). Energy balance, body composition, sedentariness and appetite regulation: pathways to obesity. *Clinical Science*, *130*(18), 1615–1628. https://doi.org/10.1042/CS20160006

Hunter, G. R., Fisher, G., Neumeier, W. H., Carter, S. J., & Plaisance, E. P. (2015). Exercise Training and Energy Expenditure following Weight Loss. *Medicine and Science in Sports and Exercise*, *47*(9), 1950–1957. https://doi.org/10.1249/MSS.0000000000000622

IAEA. (n.d.). *DLW Database*. Retrieved January 20, 2021, from https://doubly-labelled-water-database.iaea.org/dataOverview

Imboden, M. T., Nelson, M. B., Kaminsky, L. A., & Montoye, A. H. (2018). Comparison of four Fitbit and Jawbone activity monitors with a research-grade ActiGraph accelerometer for estimating physical activity and energy expenditure. *British Journal of Sports Medicine*, *52*(13), 844–850. https://doi.org/10.1136/bjsports-2016-096990

Jakicic, J. M., Marcus, B. H., Lang, W., & Janney, C. (2008). *Effect of exercise on 24-month weight loss maintenance in overweight women*. *168*(14), 1550–1559. https://doi.org/10.1001/archinte.168.14.1550

Jakicic, J. M., Marcus, M., Gallagher, K. I., Randall, C., Thomas, E., Goss, F. L., & Robertson, R. J. (2004). Evaluation of the SenseWear Pro Armband to Assess Energy Expenditure during Exercise. *Medicine & Science in Sports & Exercise*, *36*(5), 897–904. https://doi.org/10.1249/01.MSS.0000126805.32659.43

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer Texts*.

Jebb, S. A., Cole, T. J., Doman, D., Murgatroyd, P. R., & Prentice, A. M. (2000). Evaluation of the novel Tanita body-fat analyser to measure body composition by comparison with a four-compartment model. *The British Journal of Nutrition*, *83*(2), 115–122. http://www.ncbi.nlm.nih.gov/pubmed/10743490

Jefferis, B. J., Sartini, C., Shiroma, E., Whincup, P. H., Wannamethee, S. G., & Lee, I. M. (2015). Duration and breaks in sedentary behaviour: Accelerometer data from 1566 community-dwelling older men (British Regional Heart Study). *British Journal of Sports Medicine*, *49*(24), 1591–1594. https://doi.org/10.1136/bjsports-2014-093514

Jiang, L., Penney, K. L., Giovannucci, E., Kraft, P., & Wilson, K. M. (2018). A genome-wide association study of energy intake and expenditure. *PLoS ONE*, *13*(8). https://doi.org/10.1371/journal.pone.0201555

Johannsen, D. L., Calabro, M. A., Stewart, J., Franke, W., Rood, J. C., & Welk, G. J. (2010). Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. *Medicine and Science in Sports and Exercise*, *42*(11), 2134–2140. https://doi.org/10.1249/MSS.0b013e3181e0b3ff

Johansson, E., Mathiassen, S. E., Lund Rasmusse, C., & Hallman, D. M. (2020). Sitting, standing and moving during work and leisure among male and female office workers of different age: A compositional data analysis. *BMC Public Health*, *20*(1). https://doi.org/10.1186/s12889-020-08909-w

Johnson, R. K. (2002). Dietary Intake-How Do We Measure What People Are *Really* Eating? *Obesity Research*, *10*(S11), 63S-68S. https://doi.org/10.1038/oby.2002.192

Jones, A. Y. M., Kam, C., Lai, K. W., Lee, H. Y., Chow, H. T., Lau, S. F., Wong, L. M., & He, J. (2003). Changes in heart rate and R-wave amplitude with posture. *Chinese Journal of Physiology*.

Kapteyn, A., Banks, J., Hamer, M., Smith, J. P., Steptoe, A., Van Soest, A., Koster, A., & Wah, S. H. (2018). What they say and what they do: Comparing physical activity across the USA, England and the Netherlands. *Journal of Epidemiology and Community Health*, *72*(6), 471–476. https://doi.org/10.1136/jech-2017-209703

Karvonen, M. J., Kentala, E., & Mustala, O. (1957). The effects of training on heart rate; a longitudinal study. *Annales Medicinae Experimentalis et Biologiae Fenniae*, *35*(3), 307–315. http://www.ncbi.nlm.nih.gov/pubmed/13470504

Katapally, T. R., & Muhajarine, N. (2014). Towards uniform accelerometry analysis: a standardization methodology to minimize measurement bias due to systematic accelerometer wear-time variation. *Journal of Sports Science & Medicine*, *13*(2), 379–386. http://www.ncbi.nlm.nih.gov/pubmed/24790493

Kate, R. J., Swartz, A. M., Welch, W. A., & Strath, S. J. (2016). Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data. *Physiological Measurement*, *37*(3), 360–379. https://doi.org/10.1088/0967-

3334/37/3/360

Keim, N. L., Stern, J. S., & Havel, P. J. (1998). Relation between circulating leptin concentrations and appetite during a prolonged, moderate energy deficit in women. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/68.4.794

Kelvin, Lord. (1883). Electrical Units of Measurement. In *a Lecture Given on 3 May 1883, Published in the Book "Popular Lectures and Addresses, Volume 1," 1891.*

Kennedy, G. C. (1953). The role of depot fat in the hypothalamic control of food intake in the rat. *Proceedings of the Royal Society of London. Series B, Biological Sciences*. https://doi.org/10.1098/rspb.1953.0009

Kennedy, S., Ryan, L., Fraser, A., & Clegg, M. E. (2014). Comparison of the GEM and the ECAL indirect calorimeters against the Deltatrac for measures of RMR and diet-induced thermogenesis. *Journal of Nutritional Science*, *3*, e52. https://doi.org/10.1017/jns.2014.58

Kerns, J. C., Guo, J., Fothergill, E., Howard, L., Knuth, N. D., Brychta, R., Chen, K. Y., Skarulis, M. C., Walter, P. J., & Hall, K. D. (2017). Increased Physical Activity Associated with Less Weight Regain Six Years After "The Biggest Loser" Competition. *Obesity*, *25*(11), 1838–1843. https://doi.org/10.1002/oby.21986

Kerr, J., Marshall, S. J., Godbole, S., Chen, J., Legge, A., Doherty, A. R., Kelly, P., Oliver, M., Badland, H. M., & Foster, C. (2013). Using the SenseCam to improve classifications of sedentary behavior in free-living settings. *American Journal of Preventive Medicine*, *44*(3), 290–296. https://doi.org/10.1016/j.amepre.2012.11.004

Kim, D., Lee, J., Park, H. K., Jang, D. P., Song, S., Cho, B. H., Jung, Y. S., Park, R. W., Joo, N. S., & Kim, I. Y. (2017). Comparing the standards of one metabolic equivalent of task in accurately estimating physical activity energy expenditure based on acceleration. *Journal of Sports Sciences*, *35*(13), 1279–1286. https://doi.org/10.1080/02640414.2016.1221520

Kim, Y., & Welk, G. J. (2015). Criterion validity of competing accelerometry-based activity monitoring devices. *Medicine and Science in Sports and Exercise*, *47*(11), 2456–2463. https://doi.org/10.1249/MSS.0000000000000691

King, G. A., Torres, N., Potter, C., Brooks, T. J., & Coleman, K. J. (2004). Comparison of activity monitors to estimate energy cost of treadmill exercise. *Medicine and Science in Sports and Exercise*, *36*(7), 1244–1251. https://doi.org/10.1249/01.MSS.0000132379.09364.F8

Kleiber, M. (1947). Body size and metabolic rate. *Physiological Reviews*. https://doi.org/10.1152/physrev.1947.27.4.511

Klem, M. Lou, Wing, R. R., Lang, W., McGuire, M. T., & Hill, J. O. (2000). Does weight loss maintenance become easier over time? *Obesity Research*. https://doi.org/10.1038/oby.2000.54

Koehler, K., Abel, T., Wallmann-Sperlich, B., Dreuscher, A., & Anneken, V. (2015). Energy Expenditure in Adolescents With Cerebral Palsy:

Comparison of the SenseWear Armband and Indirect Calorimetry. *Journal of Physical Activity & Health*, *12*(4), 540–545. https://doi.org/10.1123/jpah.2013-0294

Koehler, K., Braun, H., De MarIes, M., Fusch, G., Fusch, C., & Schaenzer, W. (2011). Assessing energy expenditure in male endurance athletes: Validity of the sensewear armband. *Medicine and Science in Sports and Exercise*, *43*(7), 1328–1333. https://doi.org/10.1249/MSS.0b013e31820750f5

Koehler, K., & Drenowatz, C. (2017). Monitoring Energy Expenditure Using a Multi-Sensor Device-Applications and Limitations of the SenseWear Armband in Athletic Populations. *Frontiers in Physiology*, *8*, 983. https://doi.org/10.3389/fphys.2017.00983

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koorts, H., Mattocks, C., Ness, A. R., Deere, K., Blair, S. N., Pate, R. R., & Riddoch, C. (2011). The Association Between the Type, Context, and Levels of Physical Activity Amongst Adolescents. In *Journal of Physical Activity and Health* (Vol. 8, Issue 8). http://www.alspac.bris.

Kozey, S. L., Lyden, K., Howe, C. A., Staudenmayer, J. W., & Freedson, P. S. (2010). Accelerometer output and MET values of common physical activities. *Medicine and Science in Sports and Exercise*, *42*(9), 1776–1784. https://doi.org/10.1249/MSS.0b013e3181d479f2

Kraschnewski, J. L., Boan, J., Esposito, J., Sherwood, N. E., Lehman, E. B., Kephart, D. K., & Sciamanna, C. N. (2010). Long-term weight loss maintenance in the United States. *International Journal of Obesity*, *34*(11), 1644–1654. https://doi.org/10.1038/ijo.2010.94

Kräuchi, K., & Wirz-Justice, A. (2001). Circadian Clues to Sleep Onset Mechanisms. *Neuropsychopharmacology*, *25*(5), S92–S96. https://doi.org/10.1016/S0893-133X(01)00315-3

Kraus, W. E., Bhapkar, M., Huffman, K. M., Pieper, C. F., Krupa Das, S., Redman, L. M., Villareal, D. T., Rochon, J., Roberts, S. B., Ravussin, E., Holloszy, J. O., & Fontana, L. (2019). 2 years of calorie restriction and cardiometabolic risk (CALERIE): exploratory outcomes of a multicentre, phase 2, randomised controlled trial. *The Lancet Diabetes and Endocrinology*, *7*(9), 673–683. https://doi.org/10.1016/S2213-8587(19)30151-2

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Lagerpusch, M., Bosy-Westphal, A., Kehden, B., Peters, A., & Müller, M. J. (2012). Effects of brief perturbations in energy balance on indices of

glucose homeostasis in healthy lean men. *International Journal of Obesity*, *36*(8), 1094–1101. https://doi.org/10.1038/ijo.2011.211

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lam, Y. Y., & Ravussin, E. (2016). Analysis of energy metabolism in humans: A review of methodologies. *Molecular Metabolism*, *5*(11), 1057–1071. https://doi.org/10.1016/j.molmet.2016.09.005

Lam, Y. Y., & Ravussin, E. (2017). Indirect calorimetry: An indispensable tool to understand and predict obesity. *European Journal of Clinical Nutrition*, *71*(3), 318–322. https://doi.org/10.1038/ejcn.2016.220

LaMonte, M. J., Buchner, D. M., Rillamas-Sun, E., Di, C., Evenson, K. R., Bellettiere, J., Lewis, C. E., Lee, I. M., Tinker, L. F., Seguin, R., Zaslovsky, O., Eaton, C. B., Stefanick, M. L., & LaCroix, A. Z. (2018). Accelerometer-Measured Physical Activity and Mortality in Women Aged 63 to 99. *Journal of the American Geriatrics Society*. https://doi.org/10.1111/jgs.15201

Lanningham-Foster, L. M., Jensen, T. B., McCrady, S. K., Nysse, L. J., Foster, R. C., & Levine, J. A. (2005). Laboratory measurement of posture allocation and physical activity in children. *Medicine and Science in Sports and Exercise*, *37*(10), 1800–1805.

Larsen, T. M., Dalskov, S., Van Baak, M., Jebb, S., Kafatos, A., Pfeiffer, A., Martinez, J. A., Handjieva-Darlenska, T., Kunešová, M., Holst, C., Saris, W. H. M., & Astrup, A. (2010). The diet, obesity and genes (diogenes) dietary study in eight European countries - A comprehensive design for long-term intervention. *Obesity Reviews*, *11*(1), 76–91. https://doi.org/10.1111/j.1467-789X.2009.00603.x

Lee, C. M., Gorelick, M., & Mendoza, A. (2011). Accuracy of an infrared led device to measure heart rate and energy expenditure during rest and exercise. *Journal of Sports Sciences*, *29*(15), 1645–1653. https://doi.org/10.1080/02640414.2011.609899

Lee, I. M., Shiroma, E. J., Evenson, K. R., Kamada, M., LaCroix, A. Z., & Buring, J. E. (2018). Accelerometer-measured physical activity and sedentary behavior in relation to all-cause mortality: The women's health study. In *Circulation*. https://doi.org/10.1161/CIRCULATIONAHA.117.031300

Lee, I. M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., & Katzmarzyk, P. T. (2012). Impact of Physical Inactivity on the World's Major Non-Communicable Diseases. *The Lancet*.

Lee, J.-M., Kim, Y., & Welk, G. J. (2014). Validity of consumer-based physical activity monitors. *Medicine and Science in Sports and Exercise*, *46*(9), 1840–1848. https://doi.org/10.1249/MSS.0000000000000287

Lee, J. A., & Gill, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Statistical Methods in Medical Research*, *27*(2), 490–506. https://doi.org/10.1177/0962280216633248

Lee, J. M., Kim, H. C., Kang, J. I., & Suh, I. (2014). Association between stressful life events and resting heart rate. *BMC Psychology*, *2*(1), 1–9. https://doi.org/10.1186/s40359-014-0029-0

Lee, P. H. (2013). Data imputation for accelerometer-measured physical activity: The combined approach. *American Journal of Clinical Nutrition*, *97*(5), 965–971. https://doi.org/10.3945/ajcn.112.052738

Leibel, R. L., & Hirsch, J. (1984). Diminished energy requirements in reduced-obese patients. *Metabolism*. https://doi.org/10.1016/0026-0495(84)90130-6

Leibel, R. L., Rosenbaum, M., & Hirsch, J. (1995). Changes in Energy Expenditure Resulting from Altered Body Weight. *New England Journal of Medicine*, *332*(10), 621–628. https://doi.org/10.1056/NEJM199503093321001

Lejeune, M. P. G. M., Westerterp, K. R., Adam, T. C. M., Luscombe-Marsh, N. D., & Westerterp-Plantenga, M. S. (2006). Ghrelin and glucagon-like peptide 1 concentrations, 24-h satiety, and energy and substrate metabolism during a high-protein diet and measured in a respiration chamber. *American Journal of Clinical Nutrition*, *83*(1), 89–94. https://doi.org/10.1093/ajcn/83.1.89

Leonard, W. R. (2003). Measuring human energy expenditure: What have we learned from the flex-heart rate method? *American Journal of Human Biology*, *15*(4), 479–489. https://doi.org/10.1002/ajhb.10187

Levine, J. A., Eberhardt, N. L., Jensen, M. D., Levine, J. A., Eberhardt, N. L., & Jensen, M. D. (1999). *Role of Nonexercise Activity Thermogenesis in Resistance to Fat Gain in Humans Published by : American Association for the Advancement of Science Stable URL : http://www.jstor.org/stable/2897401 Linked references are available on JSTOR for this article : 283*(5399), 212–214.

Lewis, Z. H., Pritting, L., Picazo, A. L., & JeanMarie-Tucker, M. (2020). The utility of wearable fitness trackers and implications for increased engagement: An exploratory, mixed methods observational study. *Digital Health*, *6*. https://doi.org/10.1177/2055207619900059

Lindsay, T., Westgate, K., Wijndaele, K., Hollidge, S., Kerrison, N., Forouhi, N., Griffin, S., Wareham, N., & Brage, S. (2019). Descriptive epidemiology of physical activity energy expenditure in UK adults (The Fenland study). *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 126. https://doi.org/10.1186/s12966-019-0882-6

Lissner, L., Habicht, J. P., Strupp, B. J., Levitsky, D. A., Haas, J. D., & Roe, D. A. (1989). Body composition and energy intake: do overweight women overeat and underreport? *The American Journal of Clinical Nutrition*, *49*(2), 320–325. https://doi.org/10.1093/ajcn/49.2.320

Liu, B., Yu, M., Graubard, B. I., Troiano, R. P., & Schenker, N. (2016). Multiple imputation of completely missing repeated measures data within person from a complex sample: application to accelerometer data in the National Health and Nutrition Examination Survey. *Statistics in Medicine*, *35*(28), 5170–5188. https://doi.org/10.1002/sim.7049

Livesey, G., & Elia, M. (1988). Estimation of energy expenditure, net carbohydrate utilization, and net fat oxidation and synthesis by indirect calorimetry: Evaluation of errors with special reference to the detailed composition of fuels. *American Journal of Clinical Nutrition*, *47*(4), 608–628. https://doi.org/10.1093/ajcn/47.4.608

Livingston, E. H., & Kohlstadt, I. (2005). Simplified resting metabolic rate-predicting formulas for normal-sized and obese individuals. *Obesity Research*. https://doi.org/10.1038/oby.2005.149

Livingstone, M. Barbara E, Prentice, A. M., Coward, W. A., Ceesay, S. M., Strain, J. J., McKenna, P. G., Nevin, G. B., Barker, M. E., & Hickey, R. J. (1990). Simultaneous measurement of free-living energy expenditure by the doubly labeled water method and heart-rate monitoring. *American Journal of Clinical Nutrition*, *52*(1), 59–65. http://www.ncbi.nlm.nih.gov/pubmed/2193501

Livingstone, M B E, Robson, P. J., Black, A. E., Coward, W. A., Wallace, J. M. W., McKinley, M. C., Strain, J. J., & McKenna, P. G. (2003). An evaluation of the sensitivity and specificity of energy expenditure measured by heart rate and the Goldberg cut-off for energy intake: basal metabolic rate for identifying mis-reporting of energy intake by adults and children: a retrospective analysis. *European Journal of Clinical Nutrition*, *57*(3), 455–463. https://doi.org/10.1038/sj.ejcn.1601563

Lo, F. P. W., Sun, Y., Qiu, J., & Lo, B. (2020). Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review. In *IEEE Journal of Biomedical and Health Informatics* (Vol. 24, Issue 7, pp. 1926–1939). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/JBHI.2020.2987943

Lobstein, T. (2007). *Foresight Tackling Obesities: Future Choices-International Comparisons of Obesity Trends, determinants and responses-Evidence review 2 Children*. www.foresight.gov.uk

Löf, M., Henriksson, H., & Forsum, E. (2013). Evaluations of actiheart, IDEEA® and RT3 monitors for estimating activity energy expenditure in free-living women. *Journal of Nutritional Science*, *2*, 1–10. https://doi.org/10.1017/jns.2013.18

Lopez, G. A., Brønd, J. C., Andersen, L. B., Dencker, M., & Arvidsson, D. (2018). Validation of SenseWear Armband in children, adolescents, and adults. *Scandinavian Journal of Medicine and Science in Sports*, *28*(2), 487–495. https://doi.org/10.1111/sms.12920

Loprinzi, P. D., Cardinal, B. J., Crespo, C. J., Brodowicz, G. R., Andersen, R. E., & Smit, E. (2013). Differences in demographic, behavioral, and biological variables between those with valid and invalid accelerometry data: Implications for generalizability. *Journal of Physical Activity and Health*, *10*(1), 79–84. https://doi.org/10.1123/jpah.10.1.79

Lu, K., Yang, L., Seoane, F., Abtahi, F., Forsman, M., Lindecrantz, K., Ke, L., Yang, L., Seoane, F., Abtahi, F., Forsman, M., & Lindecrantz, K. (2018). Fusion of heart rate, respiration and motion measurements from a wearable sensor system to enhance energy expenditure estimation.

*Sensors (Switzerland)*, *18*(9), 3092. https://doi.org/10.3390/s18093092

Lyden, K., Keadle, S. K., Staudenmayer, J., & Freedson, P. S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine and Science in Sports and Exercise*, *46*(2), 386–397. https://doi.org/10.1249/MSS.0b013e3182a42a2d

Lyden, K., Kozey, S. L., Staudenmeyer, J. W., & Freedson, P. S. (2011). A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations. *European Journal of Applied Physiology*, *111*(2), 187–201. https://doi.org/10.1007/s00421-010-1639-8

Lyons, E. J., Lewis, Z. H., Mayrsohn, B. G., & Rowland, J. L. (2014). Behavior change techniques implemented in electronic lifestyle activity monitors: A systematic content analysis. *Journal of Medical Internet Research*, *16*(8), e192. https://doi.org/10.2196/jmir.3469

Macdiarmid, J., & Blundell, J. (1998). Assessing dietary intake: Who, what and why of under-reporting. *Nutrition Research Reviews*, *11*(2), 231–253. https://doi.org/10.1079/NRR19980017

MacDonald, H. V., Johnson, B. T., Huedo-Medina, T. B., Livingston, J., Forsyth, K. C., Kraemer, W. J., Farinatti, P. T. V., & Pescatello, L. S. (2016). Dynamic resistance training as stand-alone antihypertensive lifestyle therapy: A meta-analysis. *Journal of the American Heart Association*, *5*(10). https://doi.org/10.1161/JAHA.116.003231

Machač, S., Procházka, M., Radvanský, J., & Slabý, K. (2013). Validation of Physical Activity Monitors in Individuals with Diabetes: Energy Expenditure Estimation by the Multisensor SenseWear Armband Pro3 and the Step Counter Omron HJ-720 Against Indirect Calorimetry During Walking. *Diabetes Technology & Therapeutics*, *15*(5), 413–418. https://doi.org/10.1089/dia.2012.0235

MacKey, D. C., Manini, T. M., Schoeller, D. A., Koster, A., Glynn, N. W., Goodpaster, B. H., Satterfield, S., Newman, A. B., Harris, T. B., & Cummings, S. R. (2011). Validation of an armband to measure daily energy expenditure in older adults. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *66 A*(10), 1108–1113. https://doi.org/10.1093/gerona/glr101

MacLean, P. S., Bergouignan, A., Cornier, M. A., & Jackman, M. R. (2011). Biology's response to dieting: The impetus for weight regain. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, *301*(3), R581–R600. https://doi.org/10.1152/ajpregu.00755.2010

Maclean, P. S., Higgins, J. A., Giles, E. D., Sherk, V. D., & Jackman, M. R. (2015). The role for adipose tissue in weight regain after weight loss. *Obesity Reviews*, *16*(S1), 45–54. https://doi.org/10.1111/obr.12255

MacLean, P. S., Wing, R. R., Davidson, T., Epstein, L., Goodpaster, B., Hall, K. D., Levin, B. E., Perri, M. G., Rolls, B. J., Rosenbaum, M., Rothman, A. J., & Ryan, D. (2015). NIH working group report: Innovative research to improve maintenance of weight loss. *Obesity*, *23*(1), 7–15. https://doi.org/10.1002/oby.20967

Macridis, S., Johnston, N., Johnson, S., & Vallance, J. K. (2018). Consumer physical activity tracking device ownership and use among a population-based sample of adults. *PLoS ONE*, *13*(1). https://doi.org/10.1371/journal.pone.0189298

Maeda, H., Cho, C. C., Cho, Y., & Strath, S. J. (2019). Comparing Methods for Using Invalid Days in Accelerometer Data to Improve Physical Activity Measurement. *Journal for the Measurement of Physical Behaviour*, *2*(1), 4–12. https://doi.org/10.1123/jmpb.2018-0015

Magkos, F., Fraterrigo, G., Yoshino, J., Luecking, C., Kirbach, K., Kelly, S. C., De Las Fuentes, L., He, S., Okunade, A. L., Patterson, B. W., & Klein, S. (2016). Effects of Moderate and Subsequent Progressive Weight Loss on Metabolic Function and Adipose Tissue Biology in Humans with Obesity. *Cell Metabolism*, *23*(4), 591–601. https://doi.org/10.1016/j.cmet.2016.02.005

Maraki, M., & Sidossis, L. S. (2010). Effects of energy balance on postprandial triacylglycerol metabolism. In *Current Opinion in Clinical Nutrition and Metabolic Care* (Vol. 13, Issue 6, pp. 608–617). https://doi.org/10.1097/MCO.0b013e32833f1aae

Marlatt, K. L., Redman, L. M., Burton, J. H., Martin, C. K., & Ravussin, E. (2017). Persistence of weight loss and acquired behaviors 2 y after stopping a 2-y calorie restriction intervention. *American Journal of Clinical Nutrition*, *105*(4), 928–935. https://doi.org/10.3945/ajcn.116.146837

Martien, S., Seghers, J., Boen, F., & Delecluse, C. (2015). Energy expenditure in institutionalized older adults: Validation of SenseWear Mini. *Medicine and Science in Sports and Exercise*, *47*(6), 1265–1271. https://doi.org/10.1249/MSS.0000000000000529

Martin, C. K., Correa, J. B., Han, H., Allen, H. R., Rood, J. C., Champagne, C. M., Gunturk, B. K., & Bray, G. A. (2012). Validity of the Remote Food Photography Method (RFPM) for estimating energy and nutrient intake in near real-time. *Obesity*, *20*(4), 891–899. https://doi.org/10.1086/498510.Parasitic

Martin, C. K., Das, S. K., Lindblad, L., Racette, S. B., McCrory, M. A., Weiss, E. P., DeLany, J. P., & Kraus, W. E. (2011). Effect of calorie restriction on the free-living physical activity levels of nonobese humans: Results of three randomized trials. *Journal of Applied Physiology*, *110*(4), 956–963. https://doi.org/10.1152/japplphysiol.00846.2009

Matthews, C. E., Chen, K. Y., Freedson, P. S., Buchowski, M. S., Beech, B. M., Pate, R. R., & Troiano, R. P. (2008). Amount of time spent in sedentary behaviors in the United States, 2003-2004. *American Journal of Epidemiology*. https://doi.org/10.1093/aje/kwm390

May, R., Dandy, G., & Maier, H. (2011). Review of Input Variable Selection Methods for Artificial Neural Networks. In *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. https://doi.org/10.5772/16004

McArdle, W. D., Katch, F. I., & Katch, V. L. (2010). *Exercise physiology : nutrition, energy, and human performance*. Lippincott Williams &

Wilkins. https://capitadiscovery.co.uk/newman-ac/items/189510

McArdle, W. D., Katch, F. I., & Katch, V. L. (2015). *Essentials of exercise physiology: Fifth edition*. Wolters Kluwer Health Adis (ESP).

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. In *Journal of Clinical Epidemiology* (Vol. 67, Issue 3, pp. 267–277). Elsevier USA. https://doi.org/10.1016/j.jclinepi.2013.08.015

Mcguire, K. A., & Ross, R. (2010). Measuring Body Composition in Adults and Children. In *Clinical Obesity in Adults and Children*. Wiley-Blackwell. https://doi.org/10.1002/9781444307627.ch2

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M., Clifford, T., Cohen, J. F., Deeks, J. J., Gatsonis, C., Hooft, L., Hunt, H. A., Hyde, C. J., Korevaar, D. A., Leeflang, M. M. G., Macaskill, P., Reitsma, J. B., Rodin, R., Rutjes, A. W. S., Salameh, J.-P., … Willis, B. H. (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *JAMA*, *319*(4), 388. https://doi.org/10.1001/jama.2017.19163

McLean, J. A., & Tobin, G. (1988). Animal and Human Calorimetry. In *Animal and Human Calorimetry*. https://doi.org/10.1017/cbo9780511663161

McMinn, D., Rowe, D. A., Murtagh, S., & Nelson, N. M. (2012). The effect of a school-based active commuting intervention on children's commuting physical activity and daily physical activity. *Preventive Medicine*, *54*(5), 316–318. https://doi.org/10.1016/j.ypmed.2012.02.013

Medeiros, D. M., & Wildman, R. E. C. (2018). Advanced human nutrition. In *Jones and Bartlett Publishers, Inc*. https://doi.org/10.5860/choice.37-3933

Melanson, E. L., Dykstra, J. C., & Szuminsky, N. (2009). A novel approach for measuring energy expenditure in free-living humans. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *2009*, 6873–6877. https://doi.org/10.1109/IEMBS.2009.5333124

Melby, C. L., Paris, H. L., Foright, R. M., & Peth, J. (2017). Attenuating the biologic drive for weight regain following weight loss: Must what goes down always go back up? *Nutrients*, *9*(5). https://doi.org/10.3390/nu9050468

Mendez, M. A., Wynter, S., Wilks, R., & Forrester, T. (2004). Under- and overreporting of energy is related to obesity, lifestyle factors and food group intakes in Jamaican adults. *Public Health Nutrition*, *7*(1), 9–19.

Meng, Y., Speier, W., Shufelt, C., Joung, S., E Van Eyk, J., Bairey Merz, C. N., Lopez, M., Spiegel, B., & Arnold, C. W. (2020). A Machine Learning Approach to Classifying Self-Reported Health Status in a Cohort of Patients with Heart Disease Using Activity Tracker Data. *IEEE Journal of Biomedical and Health Informatics*, *24*(3), 878–884. https://doi.org/10.1109/JBHI.2019.2922178

Merchant, A. T., Dehghan, M., & Akhtar-Danesh, N. (2007). Seasonal variation in leisure-time physical activity among Canadians. *Canadian Journal of Public Health*, *98*(3), 203–208. https://doi.org/10.1007/bf03403713

Miller, S., Milliron, B. J., & Woolf, K. (2013). Common prediction equations overestimate measured resting metabolic rate in young hispanic women. In *Topics in Clinical Nutrition* (Vol. 28, Issue 2, pp. 120–135). NIH Public Access. https://doi.org/10.1097/TIN.0b013e31828d7a1b

Montoye, A. H. K., Begum, M., Henning, Z., & Pfeiffer, K. A. (2017). Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiological Measurement*, *38*(2), 343–357. https://doi.org/10.1088/1361-6579/38/2/343

Montoye, A. H. K., Conger, S. A., Connolly, C. P., Imboden, M. T., Nelson, M. B., Bock, J. M., & Kaminsky, L. A. (2017). Validation of Accelerometer-Based Energy Expenditure Prediction Models in Structured and Simulated Free-Living Settings. *Measurement in Physical Education and Exercise Science*, *21*(4), 223–234. https://doi.org/10.1080/1091367X.2017.1337638

Montoye, A. H. K., Mitrzyk, J. R., & Molesky, M. J. (2017). Comparative Accuracy of a Wrist-Worn Activity Tracker and a Smart Shirt for Physical Activity Assessment. *Measurement in Physical Education and Exercise Science*, *21*(4), 201–211. https://doi.org/10.1080/1091367X.2017.1331166

Montoye, A. H. K., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2015). Energy expenditure prediction using raw accelerometer data in simulated free living. *Medicine and Science in Sports and Exercise*, *47*(8), 1735–1746. https://doi.org/10.1249/MSS.0000000000000597

Montoye, A. H. K., Pivarnik, J. M., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2016). Comparison of Activity Type Classification Accuracy from Accelerometers Worn on the Hip, Wrists, and Thigh in Young, Apparently Healthy Adults. *Measurement in Physical Education and Exercise Science*, *20*(3), 173–183. https://doi.org/10.1080/1091367X.2016.1192038

Montoye, A. H. K., Westgate, B. S., Fonley, M. R., & Pfeiffer, K. A. (2018). Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer. *Journal of Applied Physiology*, *124*(5), 1284–1293. https://doi.org/10.1152/japplphysiol.00760.2017

Montoye, H. J., Kemper, H. C. G., Saris, W. H. M., & Washburn, R. A. (1996). Measuring Physical Activity and Energy Expenditure. In *Human k* (Vol. 28, Issue 9). Human Kinetics. https://doi.org/10.1097/00005768-199609000-00022

Morgan, D. W., Martin, P. E., & Krahenbuhl, G. S. (1989). Factors affecting running economy. *Sports Medicine (Auckland, N.Z.)*, *7*(5), 310–330. https://doi.org/10.2165/00007256-198907050-00003

Moshfegh, A. J., Rhodes, D. G., Baer, D. J., Murayi, T., Clemens, J. C., Rumpler, W. V., Paul, D. R., Sebastian, R. S., Kuczynski, K. J.,

Ingwersen, L. A., Staples, R. C., & Cleveland, L. E. (2008). The US Department of Agriculture Automated Multiple-Pass Method reduces bias in the collection of energy intakes. *American Journal of Clinical Nutrition*, *88*(2), 324–332. https://doi.org/10.1093/ajcn/88.2.324

Most, J., Vallo, P. M., Altazan, A. D., Gilmore, L. A., Sutton, E. F., Cain, L. E., Burton, J. H., Martin, C. K., & Redman, L. M. (2018). Food photography is not an accurate measure of energy intake in obese, pregnant women. *Journal of Nutrition*, *148*(4), 658–663. https://doi.org/10.1093/jn/nxy009

Mtaweh, H., Tuira, L., Floh, A. A., & Parshuram, C. S. (2018). Indirect calorimetry: History, technology, and application. In *Frontiers in Pediatrics* (Vol. 6, p. 257). Frontiers Media S.A. https://doi.org/10.3389/fped.2018.00257

Müller, M., Bosy-Westphal, A., & Heymsfield, S. B. (2010). Is there evidence for a set point that regulates human body weight? *F1000 Medicine Reports*, *2*, 59. https://doi.org/10.3410/M2-59

Müller, M. J., & Bosy-Westphal, A. (2013). Adaptive thermogenesis with weight loss in humans. *Obesity*, *21*(2), 218–228. https://doi.org/10.1002/oby.20027

Müller, M. J., Bosy-Westphal, A., Klaus, S., Kreymann, G., Lührmann, P. M., Neuhäuser-Berthold, M., Noack, R., Pirke, K. M., Platte, P., Selberg, O., & Steiniger, J. (2004). World Health Organization equations have shortcomings for predicting resting energy expenditure in persons from a modern, affluent population: generation of a new reference standard from a retrospective analysis of a German database of resting energy expe. *The American Journal of Clinical Nutrition*, *80*(5), 1379–1390. https://doi.org/10.1093/ajcn/80.5.1379

Müller, M. J., Geisler, C., Blundell, J., Dulloo, A., Schutz, Y., Krawczak, M., Bosy-Westphal, A., Enderle, J., & Heymsfield, S. B. (2018). The case of GWAS of obesity: does body weight control play by the rules? In *International Journal of Obesity* (Vol. 42, Issue 8, pp. 1395–1405). https://doi.org/10.1038/s41366-018-0081-6

Müller, M. J., Geisler, C., Heymsfield, S. B., & Bosy-Westphal, A. (2018). Recent advances in understanding body weight homeostasis in humans. *F1000Research*, *7*(0), 1025. https://doi.org/10.12688/f1000research.14151.1

Murakami, H., Kawakami, R., Nakae, S., Nakata, Y., Ishikawa-Takata, K., Tanaka, S., & Miyachi, M. (2016). Accuracy of wearable devices for estimating total energy expenditure: Comparisonwith metabolic chamber and doubly labeledwater method. *JAMA Internal Medicine*, *176*(5), 702–703. https://doi.org/10.1001/jamainternmed.2016.0152

Murakami, K., & Livingstone, M. B. E. (2015). Prevalence and characteristics of misreporting of energy intake in US adults: NHANES 2003-2012. *British Journal of Nutrition*, *114*(8), 1294–1303. https://doi.org/10.1017/S0007114515002706

Myers, A., Dalton, M., Gibbons, C., Finlayson, G., & Blundell, J. (2019). Structured, aerobic exercise reduces fat mass and is partially

compensated through energy intake but not energy expenditure in women. *Physiology and Behavior*, *199*, 56–65. https://doi.org/10.1016/j.physbeh.2018.11.005

Nagy, K. A. (1980). CO2 production in animals: analysis of potential errors in the doubly labeled water method. *The American Journal of Physiology*, *238*(5). https://doi.org/10.1152/ajpregu.1980.238.5.r466

Nelson, B. W., & Allen, N. B. (2019). Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. *Journal of Medical Internet Research*, *21*(3), e10828. https://doi.org/10.2196/10828

Nelson, M. B., Kaminsky, L. A., Dickin, D. C., & Montoye, A. H. K. (2016). Validity of Consumer-Based Physical Activity Monitors for Specific Activity Types. *Medicine and Science in Sports and Exercise*, *48*(8), 1619–1628. https://doi.org/10.1249/MSS.0000000000000933

Nymo, S., Coutinho, S. R., Eknes, P. H., Vestbostad, I., Rehfeld, J. F., Truby, H., Kulseng, B., & Martins, C. (2018). Investigation of the long-term sustainability of changes in appetite after weight loss. *International Journal of Obesity*, 1–11. https://doi.org/10.1038/s41366-018-0119-9

O'Driscoll, R., Turicchi, J., Beaulieu, K., Scott, S., Matu, J., Deighton, K., Finlayson, G., & Stubbs, J. (2020). How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. *British Journal of Sports Medicine*, *54*(6), 332–340. https://doi.org/10.1136/bjsports-2018-099643

O'Driscoll, R., Turicchi, J., Duarte, C., Michalowska, J., Larsen, S. C., Palmeira, A. L., Heitmann, B. L., Horgan, G. W., & Stubbs, R. J. (2020). A novel scaling methodology to reduce the biases associated with missing data from commercial activity monitors. *PLoS ONE*, *15*(6), e0235144. https://doi.org/10.1371/journal.pone.0235144

O'Driscoll, R., Turicchi, J., Hopkins, M., Gibbons, C., Larsen, S. C., Palmeira, A. L., Heitmann, B. L., Horgan, G. W., Finlayson, G., & Stubbs, R. J. (2020). The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours. *Health and Technology*, *10*(3), 637–648. https://doi.org/10.1007/s12553-019-00392-7

O'Driscoll, R., Turicchi, J., Hopkins, M., Horgan, G. W., Finlayson, G., & Stubbs, R. J. (2020). Improving energy expenditure estimates from wearable devices: A machine learning approach. *Journal of Sports Sciences*, *38*(13), 1496–1505. https://doi.org/10.1080/02640414.2020.1746088

Ostendorf, D. M., Caldwell, A. E., Creasy, S. A., Pan, Z., Lyden, K., Bergouignan, A., MacLean, P. S., Wyatt, H. R., Hill, J. O., Melanson, E. L., & Catenacci, V. A. (2019). Physical Activity Energy Expenditure and Total Daily Energy Expenditure in Successful Weight Loss Maintainers. *Obesity*, *27*(3), 496–504. https://doi.org/10.1002/oby.22373

Ostendorf, D. M., Lyden, K., Pan, Z., Wyatt, H. R., Hill, J. O., Melanson, E. L., & Catenacci, V. A. (2018). Objectively Measured Physical Activity and Sedentary Behavior in Successful Weight Loss Maintainers.

*Obesity*, *26*(1), 53–60. https://doi.org/10.1002/oby.22052

Overstreet, B. S., Bassett, D. R., Crouter, S. E., Rider, B. C., & Parr, B. B. (2017). Portable open-circuit spirometry systems. In *Journal of Sports Medicine and Physical Fitness* (Vol. 57, Issue 3, pp. 227–237). https://doi.org/10.23736/S0022-4707.16.06049-7

Pace, N., & Rathbun, E. (1945). Studies on body composition III: the body water and chemically combined nitrogen content in relation to fat content. *Journal of Biological Chemistry*. https://doi.org/10.1117/12.2046164

Papazoglou, D., Augello, G., Tagliaferri, M., Savia, G., Marzullo, P., Maltezos, E., & Liuzzi, A. (2006). Evaluation of a multisensor armband in estimating energy expenditure in obese individuals. *Obesity*, *14*(12), 2217–2223. https://doi.org/10.1038/oby.2006.260

Paraschiakos, S., de Sá, C. R., Okai, J., Slagboom, E. P., Beekman, M., & Knobbe, A. (2020). *RNNs on Monitoring Physical Activity Energy Expenditure in Older People*. http://arxiv.org/abs/2006.01169

Park, J., Kazuko, I. T., Kim, E., Kim, J., & Yoon, J. (2014). Estimating free-living human energy expenditure: Practical aspects of the doubly labeled water method and its applications. *Nutrition Research and Practice*, *8*(3), 241–248. https://doi.org/10.4162/nrp.2014.8.3.241

Pasquet, P., Brigant, L., Froment, A., Koppert, G. A., Bard, D., De Garine, I., & Apfelbaum, M. (1992). Massive overfeeding and energy balance in men: The Guru Walla model. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/56.3.483

Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490

Perez-Suarez, I., Martin-Rincon, M., Gonzalez-Henriquez, J. J., Fezzardi, C., Perez-Regalado, S., Galvan-Alvarez, V., Juan-Habib, J. W., Morales-Alamo, D., & Calbet, J. A. L. (2018). Accuracy and Precision of the COSMED K5 Portable Analyser. *Frontiers in Physiology*, *9*, 1764. https://doi.org/10.3389/fphys.2018.01764

Péronnet, F., & Massicotte, D. (1991). Table of nonprotein respiratory quotient: an update. *Canadian Journal of Sport Sciences = Journal Canadien Des Sciences Du Sport*, *16*(1), 23–29. http://www.ncbi.nlm.nih.gov/pubmed/1645211

Peyrot, N., Thivel, D., Isacco, L., Morin, J. B., Belli, A., & Duche, P. (2012). Why does walking economy improve after weight loss in obese adolescents? *Medicine and Science in Sports and Exercise*, *44*(4), 659–665. https://doi.org/10.1249/MSS.0b013e318236edd8

Pfrimer, K., Vilela, M., Resende, C. M., Scagliusi, F. B., Marchini, J. S., Lima, N. K. C., Moriguti, J. C., & Ferriolli, E. (2015). Under-reporting of

food intake and body fatness in independent older people: a doubly labelled water study. *Age and Ageing*, *44*(1), 103–108. https://doi.org/10.1093/ageing/afu142

Physical Activity Guidelines Advisory Committee. (2018). Physical Activity Guidelines Advisory Committee Scientific Report. In *United States Department of Health & Human Services*.

Plasqui, G. (2017). Smart approaches for assessing free-living energy expenditure following identification of types of physical activity. *Obesity Reviews*, *18*, 50–55. https://doi.org/10.1111/obr.12506

Pober, D. M., Staudenmayer, J., Raphael, C., & Freedson, P. S. (2006). Development of novel techniques to classify physical activity mode using accelerometers. *Medicine and Science in Sports and Exercise*, *38*(9), 1626–1634. https://doi.org/10.1249/01.mss.0000227542.43669.45

Polidori, D., Sanghvi, A., Seeley, R. J., & Hall, K. D. (2016). How Strongly Does Appetite Counter Weight Loss? Quantification of the Feedback Control of Human Energy Intake. *Obesity*, *24*(11), 2289–2295. https://doi.org/10.1002/oby.21653

Pollard, T. M., & Wagnild, J. M. (2017). Gender differences in walking (for leisure, transport and in total) across adult life: a systematic review. *BMC Public Health*, *17*(1), 341. https://doi.org/10.1186/s12889-017-4253-4

Ponce, D., Goes, C. R. de, & Andrade, L. G. M. de. (2020). Proposal of a New Equation for Estimating Resting Energy Expenditure for Acute Kidney Injury Patients on Dialysis. A Machine Learning Approach. *Nutrition & Metabolism*, *17*(1), 96. https://doi.org/10.21203/rs.3.rs-37485/v1

Pontzer, H. (2017). Economy and Endurance in Human Evolution. In *Current Biology* (Vol. 27, Issue 12, pp. R613–R621). Cell Press. https://doi.org/10.1016/j.cub.2017.05.031

Pontzer, H., Raichlen, D. A., Wood, B. M., Mabulla, A. Z. P., Racette, S. B., & Marlowe, F. W. (2012). Hunter-gatherer energetics and human obesity. *PLoS ONE*, *7*(7), e40503. https://doi.org/10.1371/journal.pone.0040503

Posma, J. M., Garcia-Perez, I., Frost, G., Aljuraiban, G. S., Chan, Q., Van Horn, L., Daviglus, M., Stamler, J., Holmes, E., Elliott, P., & Nicholson, J. K. (2020). Nutriome–metabolome relationships provide insights into dietary intake and metabolism. *Nature Food*, *1*(7), 426–436. https://doi.org/10.1038/s43016-020-0093-y

Prentice, A. M., Black, A. E., Coward, W. A., & Cole, T. J. (1996). Energy expenditure in overweight and obese adults in affluent societies: An analysis of 319 doubly-labelled water measurements. *European Journal of Clinical Nutrition*, *50*(2), 93–97.

Price, K., Bird, S. R., Lythgo, N., Raj, I. S., Wong, J. Y. L., & Lynch, C. (2017). Validation of the Fitbit One, Garmin Vivofit and Jawbone UP activity tracker in estimation of energy expenditure during treadmill

walking and running. *Journal of Medical Engineering and Technology*, *41*(3), 208–215. https://doi.org/10.1080/03091902.2016.1253795

Racette, S. B., Das, S. K., Bhapkar, M., Hadley, E. C., Roberts, S. B., Ravussin, E., Pieper, C., DeLany, J. P., Kraus, W. E., Rochon, J., & Redman, L. M. (2012). Approaches for quantifying energy intake and %calorie restriction during calorie restriction interventions in humans: The multicenter CALERIE study. *American Journal of Physiology - Endocrinology and Metabolism*, *302*(4), E441–E448. https://doi.org/10.1152/ajpendo.00290.2011

Rao, Z. Y., Wu, X. T., Liang, B. M., Wang, M. Y., & Hu, W. (2012). Comparison of five equations for estimating resting energy expenditure in Chinese young, normal weight healthy adults. *European Journal of Medical Research*, *17*(1), 26. https://doi.org/10.1186/2047-783X-17-26

Rasmussen, L. B., Matthiessen, J., Biltoft-Jensen, A., & Tetens, I. (2007). Characteristics of misreporters of dietary intake and physical activity. *Public Health Nutrition*, *10*(3), 230–237. https://doi.org/10.1017/S136898000724666X

Ravussin, E., Harper, I. T., Rising, R., & Bogardus, C. (1991). Energy expenditure by doubly labeled water: validation in lean and obese subjects. *The American Journal of Physiology*, *261*(3 Pt 1), E402-9. http://www.ncbi.nlm.nih.gov/pubmed/1909495

Ravussin, E., Lillioja, S., Anderson, T. E., Christin, L., & Bogardus, C. (1986). Determinants of 24-hour energy expenditure in man. Methods and results using a respiratory chamber. *Journal of Clinical Investigation*, *78*(6), 1568–1578. https://doi.org/10.1172/JCI112749

Reddy, R. K., Pooni, R., Zaharieva, D. P., Senf, B., El Youssef, J., Dassau, E., Doyle Iii, F. J., Clements, M. A., Rickels, M. R., Patton, S. R., Castle, J. R., Riddell, M. C., & Jacobs, P. G. (2018). Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study. *JMIR MHealth and UHealth*, *6*(12), e10338. https://doi.org/10.2196/10338

Reece, J. D., Barry, V., Fuller, D. K., & Caputo, J. (2015). Validation of the SenseWear Armband as a Measure of Sedentary Behavior and Light Activity. *Journal of Physical Activity and Health*, *12*(9), 1229–1237. https://doi.org/10.1123/jpah.2014-0136

Reed, G. W., & Hill, J. O. (1996). Measuring the thermic effect of food. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/63.2.164

Reeve, M. D., Pumpa, K. L., & Ball, N. (2014). Accuracy of the SenseWear Armband Mini and the BodyMedia FIT in resistance training. *Journal of Science and Medicine in Sport*, *17*(6), 630–634. https://doi.org/10.1016/j.jsams.2013.08.007

Rennie, K. L., Hennings, S. J., Mitchell, J., & Wareham, N. J. (2001). Estimating energy expenditure by heart-rate monitoring without individual calibration. *Medicine and Science in Sports and Exercise*, *33*(6), 939–945. https://doi.org/10.1097/00005768-200106000-00013

Rickman, A. D., Williamson, D. A., Martin, C. K., Gilhooly, C. H., Stein, R. I., Bales, C. W., Roberts, S., & Das, S. K. (2011). The CALERIE Study: Design and methods of an innovative 25% caloric restriction intervention. *Contemporary Clinical Trials*, *32*(6), 874–881. https://doi.org/10.1016/j.cct.2011.07.002

Ridgers, N. D., & Fairclough, S. (2011). Assessing free-living physical activity using accelerometry: Practical issues for researchers and practitioners. *European Journal of Sport Science*, *11*(3), 205–213. https://doi.org/10.1080/17461391.2010.501116

Ries, D., Carriquiry, A., & Shook, R. (2018). Modeling energy balance while correcting for measurement error via free knot splines. *PLoS ONE*, *13*(8), 1–22. https://doi.org/10.1371/journal.pone.0201892

Ritz, C., Astrup, A., Larsen, T. M., & Hjorth, M. F. (2019). Weight loss at your fingertips: personalized nutrition with fasting glucose and insulin using a novel statistical approach. *European Journal of Clinical Nutrition*, *73*(11), 1529–1535. https://doi.org/10.1038/s41430-019-0423-z

Rohatgi. (2017). *WebPlotDigitizer - Extract data from plots, images, and maps*. https://automeris.io/WebPlotDigitizer/

Rosenbaum, M., & Leibel, R. L. (2010). Adaptive thermogenesis in humans. *International Journal of Obesity (2005)*, *34 Suppl 1*(0 1), S47-55. https://doi.org/10.1038/ijo.2010.184

Rosenberg, D., Kadokura, E. A., Bouldin, E. D., Miyawaki, C. E., Higano, C. S., & Hartzler, A. L. (2016). Acceptability of Fitbit for physical activity tracking within clinical care among men with prostate cancer. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2016*, 1050–1059. http://www.ncbi.nlm.nih.gov/pubmed/28269902%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5333209

Rothney, M. P., Neumann, M., Béziat, A., & Chen, K. Y. (2007). An artificial neural network model of energy expenditure using nonintegrated acceleration signals. *Journal of Applied Physiology*, *103*(4), 1419–1427. https://doi.org/10.1152/japplphysiol.00429.2007

Rousset, S., Fardet, A., Lacomme, P., Normand, S., Montaurier, C., Boirie, Y., & Morio, B. (2015). Comparison of total energy expenditure assessed by two devices in controlled and free-living conditions. *European Journal of Sport Science*, *15*(5), 391–399. https://doi.org/10.1080/17461391.2014.949309

Ruddick-Collins, L. C., King, N. A., Byrne, N. M., & Wood, R. E. (2013). Methodological considerations for meal-induced thermogenesis: measurement duration and reproducibility. *British Journal of Nutrition*, *110*(11), 1978–1986. https://doi.org/10.1017/S0007114513001451

Ryan, J., & Gormley, J. (2013). An evaluation of energy expenditure estimation by three activity monitors. *European Journal of Sport Science*, *13*(6), 681–688. https://doi.org/10.1080/17461391.2013.776639

Ryan, J. M., & Howes, L. G. (2002). Relations between alcohol consumption, heart rate, and heart rate variability in men. *Heart*, *88*(6), 641–642.

https://doi.org/10.1136/heart.88.6.641

SACN. (2011). Dietary Reference Values for Energy 2011. In *Scientific Advisory Committee on Nutrition (SACN)*.

Saidj, M., Menai, M., Charreire, H., Weber, C., Enaux, C., Aadahl, M., Kesse-Guyot, E., Hercberg, S., Simon, C., & Oppert, J. M. (2015). Descriptive study of sedentary behaviours in 35,444 French working adults: Cross-sectional findings from the ACTI-Cités study. *BMC Public Health*. https://doi.org/10.1186/s12889-015-1711-8

Sanchez-Delgado, G., Alcantara, J. M. A., Ortiz-Alvarez, L., Xu, H., Martinez-Tellez, B., Labayen, I., & Ruiz, J. R. (2018). Reliability of resting metabolic rate measurements in young adults: Impact of methods for data analysis. *Clinical Nutrition*, *37*(5), 1618–1624. https://doi.org/10.1016/j.clnu.2017.07.026

Sanghvi, A., Redman, L. M., Martin, C. K., Ravussin, E., & Hall, K. D. (2015). Validation of an inexpensive and accurate mathematical method to measure long-term changes in free-living energy intake. *American Journal of Clinical Nutrition*, *102*(2), 353–358. https://doi.org/10.3945/ajcn.115.111070

Santos-Lozano, A., Hernández-Vicente, A., Pérez-Isaac, R., Santín-Medeiros, F., Cristi-Montero, C., Casajús, J. A., & Garatachea, N. (2017). Is the SenseWear Armband accurate enough to quantify and estimate energy expenditure in healthy adults? *Annals of Translational Medicine*, *5*(5), 97–97. https://doi.org/10.21037/atm.2017.02.31

Santos, I., Sniehotta, F. F., Marques, M. M., Carraça, E. V, & Teixeira, P. J. (2017). Prevalence of personal weight control attempts in adults: a systematic review and meta-analysis. *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity*, *18*(1), 32–50. https://doi.org/10.1111/obr.12466

Sardinha, L. B., & Júdice, P. B. (2017). Usefulness of motion sensors to estimate energy expenditure in children and adults: A narrative review of studies using DLW. *European Journal of Clinical Nutrition*, *71*(3), 331–339. https://doi.org/10.1038/ejcn.2017.2

Sasaki, J. E., Hickey, A. M., Staudenmayer, J. W., John, D., Kent, J. A., & Freedson, P. S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine and Science in Sports and Exercise*, *48*(5), 941–949. https://doi.org/10.1249/MSS.0000000000000844

Saunders, P. U., Pyne, D. B., Telford, R. D., & Hawley, J. A. (2004). Factors affecting running economy in trained distance runners. In *Sports Medicine*. https://doi.org/10.2165/00007256-200434070-00005

Schnell, I., Potchter, O., Epstein, Y., Yaakov, Y., Hermesh, H., Brenner, S., & Tirosh, E. (2013). The effects of exposure to environmental factors on Heart Rate Variability: An ecological perspective. *Environmental Pollution*, *183*, 7–13. https://doi.org/10.1016/j.envpol.2013.02.005

Schneller, M. B., Pedersen, M. T., Gupta, N., Aadahl, M., & Holtermann, A. (2015). Validation of five minimally obstructive methods to estimate physical activity energy expenditure in young adults in semi-

standardized settings. *Sensors (Basel, Switzerland)*, *15*(3), 6133–6151. https://doi.org/10.3390/s150306133

Schoeller, D. A. (1983). Energy expenditure from doubly labeled water: Some fundamental considerations in humans. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/38.6.999

Schoeller, D. A. (1988). Measurement of energy expenditure in free-living humans by using doubly labeled water. In *Journal of Nutrition* (Vol. 118, Issue 11, pp. 1278–1289). https://doi.org/10.1093/jn/118.11.1278

Schoeller, D. A., Ravussin, E., Schutz, Y., Acheson, K. J., Baertschi, P., & Jéquier, E. (1986). Energy expenditure by doubly labeled water: Validation in humans and proposed calculation. *American Journal of Physiology - Regulatory Integrative and Comparative Physiology*, *250*(5 (19/5)), R823–R830. https://doi.org/10.1152/ajpregu.1986.250.5.r823

Schoeller, D. A., & Van Santen, E. (1982). Measurement of energy expenditure in humans by doubly labeled water method. *Journal of Applied Physiology Respiratory Environmental and Exercise Physiology*, *53*(4), 955–959. https://doi.org/10.1152/jappl.1982.53.4.955

Schoeller, S. A., Shay, K., & Kushner, R. F. (1997). How much physical activity is needed to minimize weight gain in previously obese women? *American Journal of Clinical Nutrition*, *66*(3), 551–556. https://doi.org/10.1093/ajcn/66.3.551

Schofield, K. L., Thorpe, H., & Sims, S. T. (2019). Resting metabolic rate prediction equations and the validity to assess energy deficiency in the athlete population. *Experimental Physiology*, *104*(4), 469–475. https://doi.org/10.1113/EP087512

Schofield, W. N. (1985). Predicting basal metabolic rate, new standards and review of previous work. *Human Nutrition. Clinical Nutrition*.

Schrack, J. A., Leroux, A., Fleg, J. L., Zipunnikov, V., Simonsick, E. M., Studenski, S. A., Crainiceanu, C., & Ferrucci, L. (2018). Using Heart Rate and Accelerometry to Define Quantity and Intensity of Physical Activity in Older Adults. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *73*(5), 668–675. https://doi.org/10.1093/gerona/gly029

Schutz, Y., Weinsier, R. L., & Hunter, G. R. (2001). Assessment of free-living physical activity in humans: An overview of currently available and proposed new measures. *Obesity Research*, *9*(6), 368–379. https://doi.org/10.1038/oby.2001.48

Scott, S. E., Duarte, C., Encantado, J., Evans, E. H., Harjumaa, M., Heitmann, B. L., Horgan, G. W., Larsen, S. C., Marques, M. M., Mattila, E., Matos, M., Mikkelsen, M. L., Palmeira, A. L., Pearson, B., Ramsey, L., Sainsbury, K., Santos, I., Sniehotta, F., Stalker, C., … Stubbs, R. J. (2019). The NoHoW protocol: A multicentre 2×2 factorial randomised controlled trial investigating an evidence-based digital toolkit for weight loss maintenance in European adults. *BMJ Open*, *9*(9), e029425. https://doi.org/10.1136/bmjopen-2019-029425

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical

Modeling with Python. In *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-011

Seale, J. L., Conway, J. M., & Canary, J. J. (1993). Seven-day validation of doubly labeled water method using indirect room calorimetry. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *74*(1), 402–409. https://doi.org/10.1152/jappl.1993.74.1.402

Segal, K. R., Chun, A., Coronel, P., Cruz-Noori, A., & Santos, R. (1992). Reliability of the measurement of postprandial thermogenesis in men of three levels of body fatness. *Metabolism*. https://doi.org/10.1016/0026-0495(92)90316-3

Shaffer, J. A., Diaz, K., Alcántara, C., Edmondson, D., Krupka, D. J., Chaplin, W. F., & Davidson, K. W. (2014). An inexpensive device for monitoring patients' weights via automated hovering. *Int J Cardiol*, *172*(2), 263–264. https://doi.org/10.1016/j.ijcard.2013.12.123

Sharifzadeh, M., Bagheri, M., Speakman, J. R., & Djafarian, K. (2020). Comparison of total and activity energy expenditure estimates from physical activity questionnaires and doubly labeled water: A systematic review and meta-analysis. *British Journal of Nutrition*, 1–33. https://doi.org/10.1017/S0007114520003049

Shcherbina, A., Mikael Mattsson, C., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., Wheeler, M. T., & Ashley, E. A. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of Personalized Medicine*, *7*(2), 1–12. https://doi.org/10.3390/jpm7020003

Shephard, R. J. (2017). Open-circuit respirometry: a brief historical review of the use of Douglas bags and chemical analyzers. In *European Journal of Applied Physiology* (Vol. 117, Issue 3, pp. 381–387). https://doi.org/10.1007/s00421-017-3556-6

Sheth, A., Jaimini, U., & Yip, H. Y. (2018). How Will the Internet of Things Enable Augmented Personalized Health? *IEEE Intelligent Systems*, *33*(1), 89–97. https://doi.org/10.1109/MIS.2018.012001556

Shetty, P. (2005). Energy requirements of adults. *Public Health Nutrition*, *8*(7A), 994–1009.

Shiroma, E. J., Lee, I. M., Schepps, M. A., Kamada, M., & Harris, T. B. (2019). Physical Activity Patterns and Mortality: The Weekend Warrior and Activity Bouts. *Medicine and Science in Sports and Exercise*, *51*(1), 35–40. https://doi.org/10.1249/MSS.0000000000001762

Shook, R. P., Hand, G. A., O'Connor, D. P., Thomas, D. M., Hurley, T. G., Hébert, J. R., Drenowatz, C., Welk, G. J., Carriquiry, A. L., & Blair, S. N. (2018). Energy intake derived from an energy balance equation, validated activity monitors, and dual x-ray absorptiometry can provide acceptable caloric intake data among young adults. *Journal of Nutrition*, *148*(3), 490–496. https://doi.org/10.1093/jn/nxx029

Silva, A. M., Santos, D. A., Matias, C. N., Júdice, P. B., Magalhães, J. P., Ekelund, U., & Sardinha, L. B. (2015). Accuracy of a combined heart rate and motion sensor for assessing energy expenditure in free-living

adults during a double-blind crossover caffeine trial using doubly labeled water as the reference method. *European Journal of Clinical Nutrition*, *69*(1), 20–27. https://doi.org/10.1038/ejcn.2014.51

Siri, W. E. (1956). BODY COMPOSITION FROM FLUID SPACES AND DENSITY: , ANALYSIS OF METHODS. *Adv Biol Med Phy*.

Siri, W. E. (1961). Body composition from fluid apaces and density: analysis of methods. In *Techniques for measuring Body composition*. https://doi.org/10.1021/ma00102a600

Slinde, F., Bertz, F., Winkvist, A., Ellegård, L., Olausson, H., & Brekke, H. K. (2013). Energy expenditure by multisensor armband in overweight and obese lactating women validated by doubly labeled water. *Obesity*, *21*(11), 2231–2235. https://doi.org/10.1002/oby.20363

Smith, K. M., Lanningham-Foster, L. M., Welk, G. J., & Campbell, C. G. (2012). Validity of the SenseWear® armband to predict energy expenditure in pregnant women. *Medicine and Science in Sports and Exercise*, *44*(10), 2001–2008. https://doi.org/10.1249/MSS.0b013e31825ce76f

Soric, M., Mikulic, P., Misigoj-Durakovic, M., Ruzic, L., Markovic, G., & Westerterp, K. R. (2012). Validation of the Sensewear Armband during recreational in-line skating. *European Journal of Applied Physiology*, *112*(3), 1183–1188. https://doi.org/10.1007/s00421-011-2045-6

Sparti, A., Delany, J. P., De La Bretonne, J. A., Sander, G. E., & Bray, G. A. (1997). Relationship between resting metabolic rate and the composition of the fat-free mass. *Metabolism: Clinical and Experimental*. https://doi.org/10.1016/S0026-0495(97)90222-5

Speakman, J. (1997). *Doubly-labelled water: Theory and Practice*. Springer Berlin.

Speakman, J. R. (1993). How should we calculate CO2 production in doubly labelled water studies of animals? *Functional Ecology*, *7*(6), 746–750.

Speakman, J. R. (1997). *Doubly Labeled Water Theory and Practice* (Vol. 70, Issue 3). https://books.google.com/books?hl=en&lr=&id=cP380nUP1fwC&oi=fnd&pg=PR11&ots=Zucm9iz1__&sig=Nuh8d7W2Gr-6FARxDFMP7_XCeug

Speakman, J. R. (1998). The history and theory of the doubly labeled water technique. *American Journal of Clinical Nutrition*, *68*(4). https://doi.org/10.1093/ajcn/68.4.932S

Speakman, J. R. (2007). A Nonadaptive Scenario Explaining the Genetic Predisposition to Obesity: The “Predation Release” Hypothesis. In *Cell Metabolism*. https://doi.org/10.1016/j.cmet.2007.06.004

Speakman, J. R., Levitsky, D. A., Allison, D. B., Bray, M. S., de Castro, J. M., Clegg, D. J., Clapham, J. C., Dulloo, A. G., Gruer, L., Haw, S., Hebebrand, J., Hetherington, M. M., Higgs, S., Jebb, S. A., Loos, R. J. F., Luckman, S., Luke, A., Mohammed-Ali, V., O’Rahilly, S., … Westerterp-Plantenga, M. S. (2011). Set points, settling points and some alternative models: theoretical options to understand how genes and environments combine to regulate body adiposity. *Disease Models*

*& Mechanisms*, *4*(6), 733–745. https://doi.org/10.1242/dmm.008698

Speakman, J. R., Stubbs, R. J., & Mercer, J. G. (2002). Does body mass play a role in the regulation of food intake? *Proceedings of the Nutrition Society*. https://doi.org/10.1079/pns2002194

Spurr, G. B. (1990). Physical activity and energy expenditure in undernutrition. *Progress in Food & Nutrition Science*, *14*(2–3), 139–192. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med3&NEWS=N&AN=2293242

Spurr, G. B., Prentice, A. M., Murgatroyd, P. R., Goldberg, G. R., Reina, J. C., & Christman, N. T. (1988). Energy expenditure from minute-by-minute heart-rate recording: Comparison with indirect calorimetry. *American Journal of Clinical Nutrition*, *48*(3), 552–559. https://doi.org/10.1093/ajcn/48.3.552

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*.

St-onge, M., Mignault, D., Allison, D. B., & Rabasa-Lhoret, R. (2007). Evaluation of a portable device to measure daily energy expenditure. *American Society for Nutrition*, *85*(3), 742–749. https://doi.org/10.1093/ajcn/85.3.742

Stackpool, C., Porcari, J., Mikat, R., Gillette, C., & Foster, C. (2014). The Accuracy of Various Activity Trackers in Estimating Steps Taken and Energy Expenditure. *Journal of Fitness Research*, *3*(3), 32–48.

Stahl, S. E., An, H. S., Dinkel, D. M., Noble, J. M., & Lee, J. M. (2016). How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport & Exercise Medicine*, *2*(1), e000106. https://doi.org/10.1136/bmjsem-2015-000106

Statista. (n.d.). • *Fitbit active users 2012-2019* | *Statista*. Retrieved January 20, 2021, from https://www.statista.com/statistics/472600/fitbit-active-users/

Staudenmayer, J., He, S., Hickey, A., Sasaki, J., & Freedson, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *119*(4), 396–403. https://doi.org/10.1152/japplphysiol.00026.2015

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., & Freedson, P. (2009). An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, *107*(4), 1300–1307. https://doi.org/10.1152/japplphysiol.00465.2009

Staudenmayer, J., Zhu, W., & Catellier, D. J. (2012). Statistical considerations in the analysis of accelerometry-based activity monitor data. *Medicine and Science in Sports and Exercise*, *44*(SUPPL. 1), S61–S67. https://doi.org/10.1249/MSS.0b013e3182399e0f

Stephens, S., Beyene, J., Tremblay, M. S., Faulkner, G., Pullnayegum, E., &

Feldman, B. M. (2018). Strategies for Dealing with Missing Accelerometer Data. *Rheumatic Disease Clinics of North America*, *44*(2), 317–326. https://doi.org/10.1016/j.rdc.2018.01.012

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ (Online)*, *339*(7713), 157–160. https://doi.org/10.1136/bmj.b2393

Stevens, J., Truesdale, K. P., McClain, J. E., & Cai, J. (2006). The definition of weight maintenance. In *International Journal of Obesity* (Vol. 30, Issue 3, pp. 391–399). Int J Obes (Lond). https://doi.org/10.1038/sj.ijo.0803175

Strath, S. J., Brage, S., & Ekelund, U. (2005). Integration of physiological and accelerometer data to improve physical activity assessment. *Medicine and Science in Sports and Exercise*, *37*(11 SUPPL.), S563–S571. https://doi.org/10.1249/01.mss.0000185650.68232.3f

Strath, S. J., Swartz, A. M., Bassett, J., O'Brien, W. L., King, G. A., & Ainsworth, B. E. (2000). Evaluation of heart rate as a method for assessing moderate intensity physical activity. *Medicine and Science in Sports and Exercise*, *32*(9 SUPPL.), S465-70. https://doi.org/10.1097/00005768-200009001-00005

Stroud, M. A., Ritz, P., Coward, W. A., Sawyer, M. B., Constantin-Teodosiu, D., Greenhaff, P. L., & Macdonald, I. A. (1997). Energy expenditure using isotope-labelled water (2H218O), exercise performance, skeletal muscle enzyme activities and plasma biochemical parameters in humans during 95 days of endurance exercise with inadequate energy intake. *European Journal of Applied Physiology and Occupational Physiology*, *76*(3), 243–252. https://doi.org/10.1007/s004210050243

Stubbs, R. J., Duarte, C., O'Driscoll, R., Turicchi, J., Kwasnicka, D., Sniehotta, F. F., Marques, M. M., Horgan, G., Larsen, S., Palmeira, A., Santos, I., Teixeira, P. J., Halford, J., & Lilienthal Heitmann, B. (2021). Position Statement The H2020 "NoHoW Project": A Position Statement on Behavioural Approaches to Longer-Term Weight Management The EASO Obesity Management Task Force. *Obesity Facts*, 1–13. https://doi.org/10.1159/000513042

Stubbs, R. J., Duarte, C., O'Driscoll, R., Turicchi, J., & Michalowska, J. (2019). Developing evidence-based behavioural strategies to overcome physiological resistance to weight loss in the general population. *Proceedings of the Nutrition Society*, *78*(4), 576–589. https://doi.org/10.1017/S0029665119001083

Stubbs, R. J., Hughes, D. A., Johnstone, A. M., Whybrow, S., Horgan, G. W., King, N., & Blundell, J. (2004). Rate and extent of compensatory changes in energy intake and expenditure in response to altered exercise and diet composition in humans. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *286*(2), R350–R358. https://doi.org/10.1152/ajpregu.00196.2003

Stubbs, R. J., O'Reilly, L. M., Whybrow, S., Fuller, Z., Johnstone, A. M.,

Livingstone, M. B. E., Ritz, P., & Horgan, G. W. (2014). Measuring the difference between actual and reported food intakes in the context of energy balance under laboratory conditions. *British Journal of Nutrition*, *111*(11), 2032–2043. https://doi.org/10.1017/S0007114514000154

Stubbs, R. J., & Turicchi, J. (2021). From famine to therapeutic weight loss: Hunger, psychological responses, and energy balance-related behaviors. *Obesity Reviews*, obr.13191. https://doi.org/10.1111/obr.13191

Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S., Sharbaugh, C. O., Trabulsi, J., Runswick, S., Ballard-Barbash, R., Sunshine, J., & Schatzkin, A. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study. *American Journal of Epidemiology*, *158*(1), 1–13. https://doi.org/10.1093/aje/kwg092

Sumithran, P., Prendergast, L. A., Delbridge, E., Purcell, K., Shulkes, A., Kriketos, A., & Proietto, J. (2011). Long-Term Persistence of Hormonal Adaptations to Weight Loss. *New England Journal of Medicine*, *365*(17), 1597–1604. https://doi.org/10.1056/NEJMoa1105816

Swartz, A. M., Strath, S. J., Bassett, D. R. J., O'Brien, W. L., King, G. A., & Ainsworth, B. E. (2000). Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Medicine and Science in Sports and Exercise*, *32*(9 Suppl), S450-6.

Swift, D. L., McGee, J. E., Earnest, C. P., Carlisle, E., Nygard, M., & Johannsen, N. M. (2018). The Effects of Exercise and Physical Activity on Weight Loss and Maintenance. *Progress in Cardiovascular Diseases*, *61*(2), 206–213. https://doi.org/10.1016/j.pcad.2018.07.014

Tamura, T. (2019). Wearable oxygen uptake and energy expenditure monitors. *Physiological Measurement*, *40*(8). https://doi.org/10.1088/1361-6579/ab3827

Tanaka, H., Monahan, K. D., & Seals, D. R. (2001). Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, *37*(1), 153–156. https://doi.org/10.1016/S0735-1097(00)01054-8

Tarasuk, V., & Beaton, G. H. (1991). The nature and individuality of within-subject variation in energy intake. *American Journal of Clinical Nutrition*, *54*(3), 464–470. https://doi.org/10.1093/ajcn/54.3.464

Tataranni, P. A., Harper, I. T., Snitker, S., Del Parigi, A., Vozarova, B., Bunt, J., Bogardus, C., & Ravussin, E. (2003). Body weight gain in free-living Pima Indians: Effect of energy intake vs expenditue. *International Journal of Obesity*, *27*(12), 1578–1583. https://doi.org/10.1038/sj.ijo.0802469

Tataranni, P. A., Larson, D. E., Snitker, S., & Ravussin, E. (1995). Thermic effect of food in humans: Methods and results from use of a respiratory chamber. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/61.5.1013

Taylor, V. H., Forhan, M., Vigod, S. N., McIntyre, R. S., & Morrison, K. M. (2013). The impact of obesity on quality of life. *Best Practice &*

*Research Clinical Endocrinology & Metabolism*, *27*(2), 139–146. https://doi.org/10.1016/j.beem.2013.04.004

Thomas, D. M., Martin, C. K., Heymsfield, S., Redman, L. M., Schoeller, D. A., & Levine, J. A. (2011). A simple model predicting individual weight change in humans. *Journal of Biological Dynamics*, *5*(6), 579–599. https://doi.org/10.1080/17513758.2010.508541

Thomas, D. M., Martin, C. K., Lettieri, S., Bredlau, C., Kaiser, K., Church, T., Bouchard, C., & Heymsfield, S. B. (2013). Can a weight loss of one pound a week be achieved with a 3500-kcal deficit¿ Commentary on a commonly accepted rule. *International Journal of Obesity*, *37*(12), 1611–1613. https://doi.org/10.1038/ijo.2013.51

Thomas, D. M., Martin, C. K., Redman, L. M., Heymsfield, S. B., Lettieri, S., Levine, J. A., Bouchard, C., & Schoeller, D. A. (2014). Effect of dietary adherence on the body weight plateau: A mathematical model incorporating intermittent compliance with energy intake prescription. *American Journal of Clinical Nutrition*, *100*(3), 787–795. https://doi.org/10.3945/ajcn.113.079822

Thomas, D. M., Schoeller, D. A., Redman, L. A., Martin, C. K., Levine, J. A., & Heymsfield, S. B. (2010). A computational model to determine energy intake during weight loss. *American Journal of Clinical Nutrition*, *92*(6), 1326–1331. https://doi.org/10.3945/ajcn.2010.29687

Thomas, D. M., Scioletti, M., & Heymsfield, S. B. (2019). Predictive Mathematical Models of Weight Loss. *Current Diabetes Reports*, *19*(10), 5–11. https://doi.org/10.1007/s11892-019-1207-5

Thomson, E. A., Nuss, K., Comstock, A., Reinwald, S., Blake, S., Pimentel, R. E., Tracy, B. L., & Li, K. (2019). Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *Journal of Sports Sciences*, *37*(12), 1411–1419. https://doi.org/10.1080/02640414.2018.1560644

Thraen-Borowski, K. M., Gennuso, K. P., & Cadmus-Bertram, L. (2017). Accelerometer-derived physical activity and sedentary time by cancer type in the United States. *PLoS ONE*, *12*(8). https://doi.org/10.1371/journal.pone.0182554

Thurber, C., Dugas, L. R., Ocobock, C., Carlson, B., Speakman, J. R., & Pontzer, H. (2019). Extreme events reveal an alimentary limit on sustained maximal human energy expenditure. *Science Advances*, *5*(6). https://doi.org/10.1126/sciadv.aaw0341

Tooze, J. A., Schoeller, D. A., Subar, A. F., Kipnis, V., Schatzkin, A., & Troiano, R. P. (2007). Total daily energy expenditure among middle-aged men and women: The OPEN Study. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/86.2.382

Trabulsi, J., & Schoeller, D. A. (2001). Evaluation of dietary assessment instruments against doubly labeled water, a biomarker of habitual energy intake. *American Journal of Physiology. Endocrinology and Metabolism*, *281*(5), E891-9. https://doi.org/10.1152/ajpendo.2001.281.5.E891

Trabulsi, J., Troiano, R. P., Subar, A. F., Sharbaugh, C., Kipnis, V., Schatzkin, A., & Schoeller, D. A. (2003). Precision of the doubly labeled water method in a large-scale application: Evaluation of a streamlined-dosing protocol in the Observing Protein and Energy Nutrition (OPEN) study. *European Journal of Clinical Nutrition*, *57*(11), 1370–1377. https://doi.org/10.1038/sj.ejcn.1601698

Tremmel, M., Gerdtham, U.-G., Nilsson, P. M., & Saha, S. (2017). Economic Burden of Obesity: A Systematic Literature Review. *International Journal of Environmental Research and Public Health*, *14*(4). https://doi.org/10.3390/ijerph14040435

Troiano, R. P., Berrigan, D., Dodd, K. W., Mâsse, L. C., Tilert, T., & Mcdowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, *40*(1), 181–188. https://doi.org/10.1249/mss.0b013e31815a51b3

Troiano, R. P., McClain, J. J., Brychta, R. J., & Chen, K. Y. (2014). Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine*, *48*(13), 1019–1023. https://doi.org/10.1136/bjsports-2014-093546

Tucker, W. J., Bhammar, D. M., Sawyer, B. J., Buman, M. P., & Gaesser, G. A. (2015). Validity and reliability of Nike + Fuelband for estimating physical activity energy expenditure. *BMC Sports Science, Medicine and Rehabilitation*, *7*(1), 14. https://doi.org/10.1186/s13102-015-0008-7

Tudor-Locke, C., Camhi, S. M., & Troiano, R. P. (2012). A catalog of rules, variables, and definitions applied to accelerometer data in the national health and nutrition examination Survey, 2003-2006. *Preventing Chronic Disease*, *9*(6), E113. https://doi.org/10.5888/pcd9.110332

Tudor-Locke, C., McClain, J. J., Hart, T. L., Sisson, S. B., & Washington, T. L. (2009). Pedometry methods for assessing free-living youth. *Research Quarterly for Exercise and Sport*, *80*(2), 175–184. https://doi.org/10.1080/02701367.2009.10599551

Turicchi, J., O'Driscoll, R., Horgan, G., Duarte, C., Palmeira, A. L., Larsen, S. C., Heitmann, B. L., & Stubbs, J. (2020). Weekly, seasonal and holiday body weight fluctuation patterns among individuals engaged in a European multi-centre behavioural weight loss maintenance intervention. *PLoS ONE*, *15*(4), e0232152. https://doi.org/10.1371/journal.pone.0232152

Turicchi, J., O'Driscoll, R., Horgan, G., Duarte, C., Santos, I., Encantado, J., Palmeira, A. L., Larsen, S. C., Olsen, J. K., Heitmann, B. L., & Stubbs, R. J. (2020). Body weight variability is not associated with changes in risk factors for cardiometabolic disease. *International Journal of Cardiology: Hypertension*, *6*, 100045. https://doi.org/10.1016/j.ijchy.2020.100045

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Van Hoye, K., Boen, F., & Lefevre, J. (2015). Validation of the SenseWear Armband in different ambient temperatures. *Journal of Sports Sciences*,

*33*(10), 1007–1018. https://doi.org/10.1080/02640414.2014.981846

Van Hoye, K., Mortelmans, P., & Lefevre, J. (2014). Validation of the SenseWear Pro3 armband using an incremental exercise test. *Journal of Strength and Conditioning Research*, *28*(10), 2806–2814. https://doi.org/10.1519/JSC.0b013e3182a1f836

Van Remoortel, H., Giavedoni, S., Raste, Y., Burtin, C., Louvaris, Z., Gimeno-Santos, E., Langer, D., Glendenning, A. A. A., Hopkinson, N. S. N. N. S., Vogiatzis, I., Peterson, B. T., Wilson, F. J., Mann, B., Rabinovich, R. A., Puhan, M. A., Troosters, T., Chiesi Farmaceutici, S. A., Brindicci, C., Higenbottam, T., … E., N. (2012). Validity of activity monitors in health and chronic disease: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 84. https://doi.org/10.1186/1479-5868-9-84

van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. In *Scotts Valley, CA*.

Vanhelst, J., Mikulovic, J., Bui-Xuan, G., Dieu, O., Blondeau, T., Fardy, P., & Beghin, L. (2012). Comparison of two ActiGraph accelerometer generations in the assessment of physical activity in free living conditions. *BMC Research Notes*, *5*, 187. https://doi.org/10.1186/1756-0500-5-187

Varkevisser, R. D. M., van Stralen, M. M., Kroeze, W., Ket, J. C. F., & Steenhuis, I. H. M. (2019). Determinants of weight loss maintenance: a systematic review. In *Obesity Reviews* (Vol. 20, Issue 2). https://doi.org/10.1111/obr.12772

Vernillo, G., Savoldelli, A., Pellegrini, B., & Schena, F. (2015). Validity of the SenseWear Armband to Assess Energy Expenditure in Graded Walking. *Journal of Physical Activity and Health*, *12*(2), 178–183. https://doi.org/10.1123/jpah.2013-0437

Vooijs, M., Alpay, L. L., Snoeck-Stroband, J. B., Beerthuizen, T., Siemonsma, P. C., Abbink, J. J., Sont, J. K., & Rövekamp, T. A. (2014). Validity and usability of low-cost accelerometers for internet-based self-monitoring of physical activity in patients with chronic obstructive pulmonary disease. *Journal of Medical Internet Research*, *16*(10), e14. https://doi.org/10.2196/ijmr.3056

Wahl, Y., Düking, P., Droszez, A., Wahl, P., & Mester, J. (2017). Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. *Frontiers in Physiology*, *8*(SEP), 725. https://doi.org/10.3389/fphys.2017.00725

Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U., & Coombes, J. S. (2016). Accuracy of heart rate watches: Implications for weight management. *PLoS ONE*, *11*(5), e0154420. https://doi.org/10.1371/journal.pone.0154420

Wang, Z., Heshka, S., Gallagher, D., Boozer, C. N., Kotler, D. P., & Heymsfield, S. B. (2000). Resting energy expenditure-fat-free mass relationship: New insights provided by body composition modeling. *American Journal of Physiology - Endocrinology and Metabolism*.

https://doi.org/10.1152/ajpendo.2000.279.3.e539

Wang, Z., Ying, Z., Bosy-Westphal, A., Zhang, J., Heller, M., Later, W., Heymsfield, S. B., & Müller, M. J. (2011). Evaluation of specific metabolic rates of major organs and tissues: comparison between men and women. *American Journal of Human Biology : The Official Journal of the Human Biology Council*, *23*(3), 333–338. https://doi.org/10.1002/ajhb.21137

Weir, J. B. (1949). New methods for calculating metabolic rate with special reference to protein metabolism. *The Journal of Physiology*, *109*(1–2), 1–9. https://doi.org/10.1113/jphysiol.1949.sp004363

Welk, G. (2002). *Physical activity assessments for health-related research*. Human Kinetics. https://books.google.co.uk/books?id=O9-vt1CZJp8C&pg=PR10&lpg=PR10&dq=physical+activity+assessment+for+health+related+research&source=bl&ots=OxxEu6uG4v&sig=TP04bxq6LQjUe_aJoxVEjTmLAPg&hl=en&sa=X&ved=0ahUKEwjptsL-laLXAhWrDsAKHbG8CeUQ6AEIOzAD#v=onepage&q=phy

Welk, G. J., McClain, J. J., Eisenmann, J. C., & Wickel, E. E. (2007). Field Validation of the MTI Actigraph and BodyMedia Armband Monitor Using the IDEEA Monitor. *Obesity (19307381)*, *15*(4), 918–928. http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=24897397&site=ehost-live

Welk, G., Kim, Y., Shook, R. P., Ellingson, L., & Lobelo, R. L. (2017). Validation of a Noninvasive, Disposable Activity Monitor for Clinical Applications. *Journal of Physical Activity & Health*, *14*(7), 546–551. http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=123887560&site=ehost-live

Westerterp, K. K. R., Jonge, L. de, Bray, G., Granata, G., Brandon, L., Weststrate, J., Reed, G., Hill, J., Tataranni, P., Westerterp, K. K. R., Wilson, S., Rolland, V., Segal, K., Weststrate, J., Ravussin, E., Venne, W. V. de, Weststrate, J., Raben, A., Labayen, I., … Martinez, J. (2004). Diet induced thermogenesis. *Nutrition & Metabolism*, *1*(1), 5. https://doi.org/10.1186/1743-7075-1-5

Westerterp, K. R. (2000). Control of Energy Expenditure in Humans. In *Endotext*. MDText.com, Inc. http://www.ncbi.nlm.nih.gov/pubmed/25905198

Westerterp, K. R. (2001). *Limits to sustainable human metabolic rate*.

Westerterp, K. R. (2013). Metabolic adaptations to over—and underfeeding—still a matter of debate? *European Journal of Clinical Nutrition*, *67*(5), 443–445. https://doi.org/10.1038/ejcn.2012.187

Westerterp, K. R. (2015). Measurement of energy expenditure. *Translational Research Methods for Diabetes, Obesity and Cardiometabolic Drug Development*, *8*(7A), 169–187. https://doi.org/10.1007/978-1-4471-4920-0_7

Westerterp, K. R. (2017). Doubly labelled water assessment of energy expenditure: principle, practice, and promise. *European Journal of Applied Physiology*, *117*(7), 1277–1285. https://doi.org/10.1007/s00421-

017-3641-x

Westerterp, K. R., Saris, W. H. M., Van Es, M., & Ten Hoor, F. (1986). Use of the doubly labeled water technique in humans during heavy sustained exercise. *Journal of Applied Physiology*, *61*(6), 2162–2167. https://doi.org/10.1152/jappl.1986.61.6.2162

Weststrate, J. A. (1993). Resting metabolic rate and diet-induced thermogenesis: A methodological reappraisal. *American Journal of Clinical Nutrition*. https://doi.org/10.1093/ajcn/58.5.592

Weyer, C., Snitker, S., Rising, R., Bogardus, C., & Ravussin, E. (1999). Determinants of energy expenditure and fuel utilization in man: effects of body composition, age, sex, ethnicity and glucose tolerance in 916 subjects. *International Journal of Obesity*, *23*(7), 715–722. https://doi.org/10.1038/sj.ijo.0800910

White, T., Westgate, K., Hollidge, S., Venables, M., Olivier, P., Wareham, N., & Brage, S. (2019). Estimating energy expenditure from wrist and thigh accelerometry in free-living adults: a doubly labelled water study. *International Journal of Obesity*, *43*(11), 2333–2342. https://doi.org/10.1038/s41366-019-0352-x

Whybrow, S., Hughes, D. A., Ritz, P., Johnstone, A. M., Horgan, G. W., King, N., Blundell, J. E., & Stubbs, R. J. (2008). The effect of an incremental increase in exercise on appetite, eating behaviour and energy balance in lean men and women feeding. *British Journal of Nutrition*, *100*(5), 1109–1115. https://doi.org/10.1017/S0007114508968240

Whybrow, S., Ritz, P., Horgan, G. W., & James Stubbs, R. (2013). An evaluation of the IDEEA$^{TM}$ activity monitor for estimating energy expenditure. *British Journal of Nutrition*, *109*(1), 173–183. https://doi.org/10.1017/S0007114512000645

Whybrow, S., Ritz, P., Horgan, G. W., & Stubbs, R. J. (2013). An evaluation of the IDEEATM activity monitor for estimating energy expenditure. *The British Journal of Nutrition*, *109*(1), 173–183. https://doi.org/https://dx.doi.org/10.1017/S0007114512000645

Willetts, M., Hollowell, S., Aslett, L., Holmes, C., & Doherty, A. (2018). Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-26174-1

Williams, G., & Frühbeck, G. (2009). *Obesity : science to practice*. Wiley. https://books.google.co.uk/books?hl=en&lr=&id=zFE03wY-eUAC&oi=fnd&pg=PR7&dq=obesity+science+to+practice&ots=ZkODKGWQ9j&sig=4cQ-CoAelex5D-F2wBtroMAN1bk#v=onepage&q=obesity science to practice&f=false

Wing, R. R., & Phelan, S. (2005). Long-term weight loss maintenance. *The American Journal of Clinical Nutrition*, *82*(1), 222S-225S. https://doi.org/10.1093/ajcn/82.1.222S

Wishnofsky, M. (1958). Caloric Equivalents of Gained or Lost Weight. *The American Journal of Clinical Nutrition*, *6*(5), 542–546.

https://doi.org/10.1093/ajcn/6.5.542

Wolfe, B. M., Schoeller, D. A., McCrady-Spitzer, S. K., Thomas, D. M., Sorenson, C. E., & Levine, J. A. (2018). Resting Metabolic Rate, Total Daily Energy Expenditure, and Metabolic Adaptation 6 Months and 24 Months After Bariatric Surgery. *Obesity*, *26*(5), 862–868. https://doi.org/10.1002/oby.22138

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. In *IEEE Transactions on Evolutionary Computation* (Vol. 1, Issue 1). https://doi.org/10.1109/4235.585893

Wong, W. W., Roberts, S. B., Racette, S. B., Das, S. K., Redman, L. M., Rochon, J., Bhapkar, M. V, Clarke, L. L., & Kraus, W. E. (2014). The doubly labeled water method produces highly reproducible longitudinal results in nutrition studies. *The Journal of Nutrition*, *144*(5), 777–783. https://doi.org/10.3945/jn.113.187823

Woodman, J. A., Crouter, S. E., Bassett, D. R., Fitzhugh, E. C., & Boyer, W. R. (2017). Accuracy of Consumer Monitors for Estimating Energy Expenditure and Activity Type. *Medicine and Science in Sports and Exercise*, *49*(2), 371–377. https://doi.org/10.1249/MSS.0000000000001090

World Health Organization. (1985). *Food and Agricultural Organization/World Health Organization/United Nations University. Energy and Protein Requirements*. Report of a Joint FAO/WHO/UNU Expert Consultation.

Wright, S. P., Hall Brown, T. S., Collier, S. R., & Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *312*(3), R358–R367. https://doi.org/10.1152/ajpregu.00349.2016

Yang, C. C., & Hsu, Y. L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, *10*(8), 7772–7788. https://doi.org/10.3390/s100807772

Yue Xu, S., Nelson, S., Kerr, J., Godbole, S., Patterson, R., Merchant, G., Abramson, I., Staudenmayer, J., & Natarajan, L. (2018). Statistical approaches to account for missing values in accelerometer data: Applications to modeling physical activity. *Statistical Methods in Medical Research*, *27*(4), 1168–1186. https://doi.org/10.1177/0962280216657119

Zhang, K., Pi-Sunyer, F. X., & Boozer, C. N. (2004). Improving Energy Expenditure Estimation for Physical Activity. *Medicine and Science in Sports and Exercise*, *36*(5), 883–889. https://doi.org/10.1249/01.MSS.0000126585.40962.22

Zhang, S., Rowlands, A. V., Murray, P., & Hurst, T. L. (2012). Physical activity classification using the GENEA wrist-worn accelerometer. *Medicine and Science in Sports and Exercise*, *44*(4), 742–748. https://doi.org/10.1249/MSS.0b013e31823bf95c

# List of Abbreviations

Device abbreviations specific to chapter 4

ACT, Actical;

AGT3X, Actigraph GT3X;

AW, Apple watch;

AWS2, Apple Watch series 2;

BA, Beurer AS80;

BMC, Bodymedia CORE armband;

BP, Basis Peak;

EP, Epson Pulsense;

EPUL, ePulse Personal Fitness Assistant;

FB, Fitbit Blaze;

FC, Fitbit Charge;

FC2, Fitbit Charge 2;

FCHR, Fitbit Charge HR;

FF, Fitbit Flex;

GF225, Garmin Forerunner 225;

GF920XT, Garmin Forerunner 920XT;

GVA, Garmin Vivoactive;

GVF, Garmin Vivofit;

GVS, Garmin vivosmart;

GVHR, Garmin Vivosmart HR;

JU, Jawbone UP;

JU24, Jawbone UP24;

LC, LifeChek calorie sensor;

MA, Mio Alpha;

MB, Microsoft band;

MS, Misfit Shine;

NF, Nike FuelBand;

PL, Polar Loop;

Polar AW200, Polar, AW200;

PA360, Polar, AW360;

SG, Samsung Gear S;

SWA, SenseWear Armband;

SWA p2, SenseWear Armband Pro 2;

SWA p3, SenseWear Armband Pro 3;

SWAM, SenseWear Armband Mini;

TT, TOMTOM Touch;

V, Vivago;

WP, Withings Pulse;

WPO, Withings Pulse O2

Abbreviations used consistently throughout the thesis

Actigraph, AG;

Activity energy expenditure, AEE;

Adenosine-monophosphate, AMP;

Adenosine triphosphate, ATP;

Analysis of variance, ANOVA;

Air displacement plethysmography, ADP;

Application programming interface, API;

Basal metabolic rate, BMR;

Beats per minute, BPM;

Body mass index, BMI;
Dual-energy x-ray absorptiometry, DEXA;

Diastolic blood pressure, DBP;

Dietary induced thermogenesis, DIT;

Doubly labelled water DLW;

Energy expenditure, EE;

Energy intake, EI;

Energy storage, ES;

Fitbit Charge 2, FB;

Fat mass, FM;

Fat-free mass, FFM;

Gradient boost, GB;

k-nearest neighbors, KNN;

Kilocalorie, KCAL;

Kilogram, KG;

Kilojoule, KJ;

Mean absolute error, MAE;

Bioelectrical impedance analysis; BIA

Mean absolute percentage error, MAPE

Megajoule, MJ;

Metabolic equivalent (MET);

Moderate-to-vigorous physical activity, MVPA;

National Institute of Diabetes and Digestive and Kidney Disease, NIDDK;

Neural network, NN;

Non-exercise activity thermogenesis, NEAT;

Physical activity energy expenditure, PAEE;

Physical activity level, PAL;

Random Forest, RF;

Resting metabolic rate, RMR;
Resting heart rate, RHR;

Root mean squared error, RMSE;

SenseWear Armband Mini, SWA;

Standard deviation, SD;

Standard error, SE;

Systolic blood pressure, SBP;

Support vector machine, SVM;

Total body water, TBW;

Total daily energy expenditure, TDEE;

Total energy expenditure from wearable devices (study), TEED;

Volume of carbon dioxide consumption, $\mathrm{VCO}_2$;

Volume of oxygen consumption, $\mathrm{VO}_2$;

World health organisation, WHO;

Weight gainer, WG;

Weight loser, WL;

Weight loss maintainer, WLM;

# Appendices

## Appendix 1.1 Search strategy

**Population:** Healthy adult populations (>18). Free from factors that impact physical movement.
**Intervention:** activity monitors + all research grade accelerometers (must be wearable on wrist or arm)
**Comparison:** Validated method: metabolic cart, DLW, DC, all IC systems,
**Outcome:** validity of energy expenditure (kcal/kj/met/correlation)

| | P | I | C | O |
|---|---|---|---|---|
| **Key concepts** | **ADULTS** | **ACTIVITY MONITORS** | **VALIDATED METHOD** | **ENERGY EXPENDITURE** |
| **Related terms** | | FITNESS TRACKERS (CINHAL)<br>ACCELEROMETRY (MESH)<br>ACCELEROMETER<br>AMBULATORY MONITOR*<br>FITBIT<br>ACTIVITY MONITOR | VALID*<br>COMPAR*<br>TEST | ENERGY METABOLISM (MESH)<br>CALORIES<br>ENERGY EXPENDITURE<br>CALORIC EXPENDITURE<br>TOTAL DAILY ENERGY EXPENDITURE<br>TDEE<br>AEE |
| **Terms to include in search** | | 1. Activity tracker<br>2. Activity Monitor<br>3. Health tracker<br>4. Health monitor<br>5. Fitness tracker<br>6. Fitness monitor<br>7. Physical activity tracker<br>8. Physical activity monitor<br>9. Exercise tracker<br>10. Exercise monitor<br>11. Electronic tracker<br>12. Electronic monitor<br>13. acceleromet<br>14. Step tracker | 1. Doubly labelled water<br>2. Dlw<br>3. Indirect caliomet*<br>4. Caliomet*<br>5. Direct caliomet*<br>6. Metabolic chamber<br>7. Metabolic cart<br>8. Gold standard<br>9. Criterion | Energy expenditure<br>1. Energy metabolism<br>2. Calori*<br>3. Calori* expenditure<br>4. Total energy expenditure<br>5. Activity energy expenditure<br>6. AEE<br>7. TDEE |

| | | | | |
|---|---|---|---|---|
| | | 15. Wearable | | |

(Tracker AND EE) AND Validation

1. Activity tracker
2. Activity Monitor
3. Health tracker
4. Health monitor
5. Fitness tracker
6. Fitness monitor
7. Physical activity tracker
8. Physical activity monitor
9. Exercise tracker
10. Exercise monitor
11. Electronic tracker
12. Electronic monitor
13. acceleromet
14. Step tracker
15. Wearable

**AND**

1. Energy expenditure
2. Energy metabolism
3. Calori*
4. Calori* expenditure
5. Total energy expenditure
6. Activity energy expenditure
7. AEE
8. TDEE

**AND**

1. Doubly labelled water
2. Dlw
3. Indirect caliomet*
4. Caliomet*
5. Direct caliomet*
6. Metabolic chamber
7. Metabolic cart
8. Gold standard
9. Criterion

| Database | Search | Results |
|---|---|---|
| Sport discus | ( activity tracker or activity monitor or health tracker or health monitor or | 154 |

| | | |
|---|---|---|
| | fitness tracker or fitness monitor or physical activity tracker or physical activity monitor or exercise tracker or exercise monitor or electronic tracker or electronic monitor or acceleromet* or step tracker or wearable tracker ) AND ( energy expenditure or energy metabolism or calori* or calori* expenditure or total energy expenditure or activ* energy expenditure or AEE or TDEE ) AND ( doubly labelled water or DLW or indirect caliomet* or caliomet* or direct caliomet* or metabolic chamber or metabolic cart or gold standard or criterion ) | |
| Pubmed | *((((((((((((((((activity tracker) OR activity monitor) OR health tracker) OR health monitor) OR fitness trackers) OR fitness monitor) OR physical activity tracker) OR physical activity monitor) OR exercise trained) OR exercise monitor) OR electronic trackers) OR electronic monitor) OR acceleromet*) OR step tracer) OR wearable trackers)) AND ((((((((energy expenditure) OR energy metabolism) OR calori*) OR calori* expenditure) OR total energy expenditure) OR activ* energy expenditure) OR AEE) OR tdee))) AND (((((((((doubly labelled water) OR DLW) OR indirect caliomet*) OR caliomet*) OR direct caliomet*) OR metabolic chamber) OR metabolic cart) OR gold standard) OR criterion)*. | 605 |

| | | |
|---|---|---|
| MEDLINE | ((activity tracker or activity monitor or health tracker or health monitor or fitness tracker or fitness monitor or physical activity tracker or physical activity monitor or exercise tracker or exercise monitor or electronic tracker or electronic monitor or acceleromet* or step tracker or wearable tracker).mp. AND (energy expenditure or energy metabolism or calori* or calori* expenditure or total energy expenditure or activ* energy expenditure or AEE or TDEE).mp. AND (doubly labelled water or DLW or indirect caliomet* or caliomet* or direct caliomet* or metabolic chamber or metabolic cart or gold standard or criterion).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word] | 228 |
| Psycinfo | ((activity tracker or activity monitor or health tracker or health monitor or fitness tracker or fitness monitor or physical activity tracker or physical activity monitor or exercise tracker or exercise monitor or electronic tracker or electronic monitor or acceleromet* or step tracker or wearable tracker).mp. AND (energy expenditure or energy metabolism or calori* or calori* expenditure or total energy expenditure or activ* energy expenditure or AEE or TDEE).mp. | 26 |

| | | |
|---|---|---|
| | AND (doubly labelled water or DLW or indirect caliomet* or caliomet* or direct caliomet* or metabolic chamber or metabolic cart or gold standard or criterion).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word] | |
| Embase | ((activity tracker or activity monitor or health tracker or health monitor or fitness tracker or fitness monitor or physical activity tracker or physical activity monitor or exercise tracker or exercise monitor or electronic tracker or electronic monitor or acceleromet* or step tracker or wearable tracker).mp. AND (energy expenditure or energy metabolism or calori* or calori* expenditure or total energy expenditure or activ* energy expenditure or AEE or TDEE).mp. AND (doubly labelled water or DLW or indirect caliomet* or caliomet* or direct caliomet* or metabolic chamber or metabolic cart or gold standard or criterion).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word] | 317 |
| CINHAL | ( activity tracker or activity monitor or health tracker or health monitor or fitness tracker or fitness monitor or physical activity tracker or physical | 142 |

| | | |
|---|---|---|
| | activity monitor or exercise tracker or exercise monitor or electronic tracker or electronic monitor or acceleromet* or step tracker or wearable tracker ) AND ( energy expenditure or energy metabolism or calori* or calori* expenditure or total energy expenditure or activ* energy expenditure or AEE or TDEE ) AND ( doubly labelled water or DLW or indirect caliomet* or caliomet* or direct caliomet* or metabolic chamber or metabolic cart or gold standard or criterion ) | |
| Obtained from reference lists | | 63<br>AFTER REMOVAL OF DUPLICATES: 825 |

**Exclusions:**

**1 = not comparison to criterion**

**2 = not comparison to accelerometer**

**3 = not healthy adult population**

**4 = review**

**5 = not kcal/kj**

**6= duplicate**

# Appendix 1.2 Study systematic review

| | Sample characteristics | Study protocol | Setting (Lab/ Field) | Criterion comparison | Device | Device placement | Results (overall error relative to criterion) |
|---|---|---|---|---|---|---|---|
| Alsubheen, 2016 | N=13 (5 F)<br>Age: 40 ± 11.9 y<br>BMI: 27 ± 4.3 kg/m$^2$ | Subjects performed a graded treadmill test. | Lab | IC – Sable system (Sable Systems International, Las Vegas NV) | Garmin vivofit (Garmin ltd, Olathe, Kansas, USA) | Wrist | Garmin vivofit:<br>-41.63% |
| Bai, 2017 | N=39 (16 F)<br>Age: 32 ± 11 y<br>BMI: 24.7 ± 4 kg/m$^2$ | Subjects performed a semi-structured activity protocol consisting of sedentary activity, aerobic exercise, and light intensity physical activity on a treadmill. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | Apple watch (Apple Inc, Cupertino, California, USA)<br>Fitbit charge HR (Fitbit Inc, San Francisco, California, USA) | Wrist | Apple Watch:<br>-10.79%<br>Fitbit Charge HR:<br>17.88% |
| Benito, 2012 | N=29 (17 F)<br>Age: 22.5 y | Subjects performed circuits of resistance exercise at 30%, 50% and 70% of 15 | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, | SenseWear Pro2 Armband (HealthWear, Bodymedia, | Upper arm | SenseWear Pro2 Armband: -46.60% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | BMI: 22 kg/m$^2$ | repetition maximum. | | Viasys Healthcare, Germany) | Pittsburgh, PA, USA) | | |
| Berntsen, 2010 | N=20 (6 F)<br>Age: 35 y<br>BMI: 24 kg/m$^2$ | Subjects performed lifestyle and sporting activities including strength exercises, ball games, occupational and home-based activities. | Lab | IC – MetaMax II (Cortex Biophysic, Leipzig, Germany) | SenseWear Pro2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro2 Armband: -9.00% |
| Berntsen, 2011 | N=29 (29 F)<br>Age: 31 ± 4.1 y<br>BMI: 27 ± 3.2 kg/m$^2$ | Subjects participated in a period of sedentary behaviour. 9 subjects then performed callisthenics and cycling on a bicycle ergometer. The other 20 subjects performed outdoor walking followed by relaxing, cycling and callisthenics. | Lab | IC – MetaMax II (Cortex Biophysic, Leipzig, Germany) | SenseWear Pro2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro2 Armband: -10.34% |
| Bhammar, 2016 | N=34 (26 F)<br>Age: 30.1 ± 8.7 y<br>BMI: 26.2 ± 5.1 kg/m$^2$ | Subjects performed a semi structured and a structured routine.<br>Semi-structured: 12 activities including 4 sedentary/light-intensity activities, 4 moderate-intensity activities, and 4 vigorous-intensity activities. The activities performed were randomly selected from a list of common activities. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Mini Armband: 14.76% |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Structured: A period of rest, followed by 7 activities of 8 minutes each. The activities performed were randomly selected from a list of common activities. | | | | |
| Boudreaux, 2018 | N=50 (28 F)<br>Age: 22.4 y<br>BMI: 26.5 kg/m$^2$ | Subjects performed separate trials of graded cycling and 3 sets of 4 resistance exercises at a 10-repetition maximum load. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | Apple Watch 2 (Apple Inc, Cupertino, California, USA)<br>Fitbit Blaze (Fitbit Inc, San Francisco, California, USA)<br>Fitbit Charge 2 (Fitbit Inc, San Francisco, California, USA)<br>Garmin Vivosmart HR (Garmin ltd, Olathe, Kansas, USA)<br>Polar: the Activity Watch 360 | Apple Watch 2: 48.20%<br>Fitbit Blaze: 28.66%<br>Fitbit Charge 2: -30.97%<br>Garmin Vivosmart HR: 16.85%<br>Polar: the Activity Watch 360: |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | (Polar Electro Oy, Kempele, Finland)<br><br>Tomtom touch (TomTom, Amsterdam, the Netherlands) | | 28.68%<br><br>Tomtom Touch: 28.66% |
| Brazeau, 2011 | N=31 (16 F)<br>Age: 26.7 y<br>BMI: 27.5 kg/$m^2$ | Subjects performed 45 minutes of stationary cycling at 50% $VO_{2peak}$. | Lab | IC – Ergocard exercise test station (MediSoft, Dinant, Belgium) | SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro3 Armband: -10.56% |
| Brazeau, 2014 | N=38 (18 F)<br>Age: 28.6 y<br>BMI: 23.8 kg/$m^2$ | Subjects performed 45 minutes of treadmill exercise at 40% $VO_{2peak}$ then exercised on a stationary bike ergometer for 45 minutes at 50% $VO_{2peak}$. | Lab | IC – Ergocard exercise test station (MediSoft, Dinant, Belgium) | SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro3 Armband:<br>14.94% |
| Brazeau, 2016 | N=20 (0 F)<br>Age: 26.2 ± 3.6 y<br>BMI: 23.1 ± 2.3 kg/$m^2$ | Subjects completed a field observation and a lab protocol.<br><br>Field: 7-day comparison to DLW.<br><br>Lab: Subjects performed 60 minutes rest | Lab/ Field | DLW – 7 days<br><br>IC – Ergocard exercise test station (MediSoft, Dinant, Belgium) | SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro3 Armband:<br>7.06% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | followed by treadmill exercise for 45 minutes at 22-41% $VO_{2peak}$ then stationary cycling for 45 minutes at 50% $VO_{2peak}$. | | | | | |
| Brugniaux, 2010 | N=31 (16 F)<br>Age: 42.9 y<br>BMI: 22.7 kg/m$^2$ | Subjects performed a 9.7km outdoor hike. | Field | IC – Metablograph with Hans Rudolph facemask (Hans Rudolph, Kansas City, MO, USA) | Polar: the Activity Watch 200 (Polar Electro Oy, Kempele, Finland) | Wrist | Polar: the Activity Watch 200: -13.17% |
| Calabro, 2014 | N=40 (19 F)<br>Age: 27.4 y<br>BMI: 22.8 kg/m$^2$ | Subjects performed 60 minutes of structured activities including stationary biking, walking/ running on a treadmill, road biking, elliptical exercise and stair stepping and unstructured movements. The semi-structured measurement periods were performed in 5, 10, 10, 10, and 25-minute intervals and included sitting, walking, standing, stair climbing or light movements. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA)<br><br>SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Mini Armband: 0.89%<br><br>SenseWear Pro3 Armband:<br>2.33% |
| Calabro, 2015 | N=29 (17 F)<br>Age: 68.8 ± 6.3 y | 14-day comparison to DLW. | Field | DLW – 14 days | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Mini Armband: -0.86% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | BMI: 26.3 ± 4.9 kg/m$^2$ | | | | | | |
| Casiraghi, 2013 | N=18 (11 F)<br>Age: 48.6 ± 21 y<br>BMI: 24.6 ± 2.6 kg/m$^2$ | Subjects performed a cycling protocol with three components:<br>1) Baseline where the subject sat on the cycle ergometer.<br>2) A 2-minute warm-up at 40 rpm at 40 watts.<br>3) Exercise increased to 60 rpm and intensity progressed by 7 watts/minute until exhaustion. | Lab | IC – SensorMedics Vmax 229 (SensorMedics Inc, Yorba Linda, CA, USA). | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -8.00% |
| Chowdhry, 2017 | N=30 (15 F)<br>Age: 27 ± 1.6 y<br>BMI: 23.4 ± 2.5 kg/m$^2$ | Subjects performed two components:<br>1) A protocol of 4 activities of designed to replicate daily living tasks<br>2) 4 activities of 10 minutes in duration. These activities were walking on a treadmill, walking at the same speed with shopping bags, cycling on an ergometer and jogging on the treadmill. | Lab | IC – COSMED K4b2 (COSMED, Rome, Italy) | Apple watch (Apple Inc, Cupertino, California, USA)<br>Microsoft Band (Microsoft Corporation, Redmond, Washington, USA)<br>Fitbit Charge HR (Fitbit Inc, San Francisco, California, | Wrist<br>Bodymedia core: Upper arm | Apple watch: -6.9%<br>Microsoft Band: -49.15%<br>Fitbit Charge HR: 15.49% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | USA)<br><br>Jawbone UP24 (Jawbone, San Francisco, California, USA)<br><br>Bodymedia Core (HealthWear, Bodymedia, Pittsburg, PA, USA) | | Jawbone UP24: -21.01%<br><br>Bodymedia Core: 7.98% |
| Colbert, 2011 | N=56 (45 F)<br>Age: 74.7 ± 6.5 y<br>BMI: 25.8 ± 4.2 kg/m$^2$ | 10-day comparison to DLW. | Field | DLW – 10 days | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband: 58.53% |
| Correa, 2016 | N=87 (72 F)<br>Age: 42 ± 13 y<br>BMI: 31.6 ± 4.5 kg/m$^2$ | 7-day comparison to DLW. | Field | DLW – 7 days | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm<br><br>Wrist | SenseWear Armband −416.95 kcal |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | Actical: 194.52 kcal |
| | | | | | Actical (Phillips Respironics Inc, Murrysville, PN, USA) | | |
| Diaz, 2015 | N=23 (13 F)<br>Age: N/A<br>BMI: N/A | Subjects performed a treadmill protocol consisting of walking at slow, moderate and brisk paces and jogging. | Lab | IC – Ultima CPX (Medgraphics, Saint Paul, MN, USA) | Fitbit Flex (Fitbit Inc, San Francisco, CA, USA) | Wrist | Fitbit Flex: 17.36% |
| Diaz, 2016 | N=13 (13 F)<br>Age: 32.0 ± 9.2 y<br>BMI: 24.2 ± 3.4 kg/m$^2$ | Subjects performed a treadmill protocol consisting of walking at slow, moderate and brisk paces and jogging. | Lab | IC – Ultima CPX (Medgraphics, Saint Paul, MN, USA) | Fitbit Flex (Fitbit Inc, San Francisco, CA, USA) | Wrist | Fitbit Flex: 30.27% |
| Dondzila, 2016 | N=19 (5 F)<br>Age: 24.6 ± 3.1 y<br>BMI: 28.0 ± 3.8 kg/m$^2$ | Subjects performed 5-minute stages of jogging on a treadmill at increasing velocity. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | Fitbit Charge (Fitbit Inc, San Francisco, California, USA) | Wrist | Fitbit Charge: -13.01% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dooley, 2017 | N=62 (36 F)<br>Age: 22.46 y<br>BMI: 24.86 kg/m$^2$ | Subjects performed 4 stages of treadmill exercise followed by a seated recovery period. The activity routine consisted of an unmeasured warm-up walking period and measured stages of slow, then brisk walking and jogging. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | Apple watch (Apple Inc, Cupertino, CA, USA)<br>Fitbit charge HR (Fitbit Inc, San Francisco, CA, USA)<br>Garmin Forerunner 225 (Garmin ltd, Olathe, Kansas, USA) | Wrist | Apple watch: 64.55%<br>Fitbit charge HR: 18.70%<br>Garmin Forerunner 225: 44.23% |
| Drenowatz, 2011 | N=20 (10 F)<br>Age: 24.3 y<br>BMI: N/A | Subjects performed three treadmill runs at 65, 75, and 85% $VO_{2max}$. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -32.80% |
| Erdogan, 2010 | N=43 (27 F)<br>Age: 34.9 ± 5.5 y<br>BMI: 31.2 ± 3.7 kg/m$^2$ | Subjects performed rowing exercises at 50% and 70% $VO_{2max}$ on an ergometer. | Lab | IC – COSMED K4b2 (COSMED, Rome, Italy) | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: 5.23% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Fruin, 2010 | Experiment 1: N=13 (0 F)<br>Experiment 2: N=20 (10 F)<br>Age: 20.2 ± 1 y<br>BMI: N/A | Experiment 1: Subjects performed two resting and a cycle ergometer session at 60% $VO_{2peak}$.<br>Experiment 2: Subjects completed a treadmill protocol of jogging, running and uphill running. | Lab | IC – SensorMedics Vmax 229 (SensorMedics Inc, Yorba Linda, CA, USA). | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -1.76% |
| Furlanetto, 2010 | N=30 (15 F)<br>Age: 68 ± 7 y<br>BMI: 25 ± 3 kg/$m^2$ | Subjects performed a walking protocol on a treadmill at three intensities. | Lab | IC – $VO_{2000}$ aerograph (Medgraphics, Saint Paul, MN, USA) | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -6.99% |
| Gastin, 2017 | N=26 (12 F)<br>Age: 21.3 ± 2.4 y<br>BMI: 23.2 ± 2 kg/$m^2$ | Subjects performed a protocol Involving resting periods, walking, jogging, running or a sport-simulated circuit. | Lab | IC – MetaMax 3b (Cortex Biophysic, Leipzig, Germany) | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -19.90% |
| Heiermann, 2011 | N=32 (19 F)<br>Age: 68.6 y<br>BMI: 26.4 kg/$m^2$ | Subjects were required to rest. | Lab | IC – Vmax Spectra (SensorMedics Viasys Healthcare, Bilthoven, The Netherlands) | SenseWear Pro2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro2 Armband: 10.80% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Imboden, 2017 | N=30 (15 F)<br>Age: 49.2 ± 19.2 y<br>BMI: 26.2 kg/m$^2$ | Subjects performed a semi-structured activity protocol, performing ≥12 activities for subject-selected duration and pace. Activities were selected from a list of sedentary, household activities ambulatory and cycling activities. | Lab | IC – COSMED K4b2 (COSMED, Rome, Italy) | Fitbit flex (Fitbit Inc, San Francisco, California, USA)<br><br>Jawbone UP24 (Jawbone, San Francisco, California, USA) | Wrist | Fitbit flex: -15.29%<br><br>Jawbone UP24:<br>-40.00% |
| Jakicic, 2004 | N=40 (20 F)<br>Age: 23.2 ± 3.8 y<br>BMI: 23.8 ± 3.1 kg/m$^2$ | Subjects performed 4 separate exercise protocols including treadmill walking, stair stepping, cycle ergometry, and arm ergometry. | Lab | IC – SensorMedics Vmax 229 (SensorMedics Inc, Yorba Linda, CA, USA). | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -11.76% |
| Johannsen, 2010 | N=30 (15 F)<br>Age: 38.2 ± 10.6 y<br>BMI: 24 ± 3.4 kg/m$^2$ | 14-day comparison to DLW. | Field | DLW – 14 days | SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA)<br><br>SenseWear Mini Armband HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro3 Armband: -2.48% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Kim, 2015 | N=52 (19 F)<br>Age: 23.8 ± 5.2<br>BMI: N/A | Subjects performed 15 activities including resting, stair climbing, cycling, walking and jogging. Each activity was performed for 5 minutes, with 1-minute resting intervals. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | Bodymedia Core (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | Bodymedia Core: 5.80% |
| King, 2004 | N=21 (10 F)<br>Age: 37.55 y | Subjects performed 10 minutes of treadmill walking and running at various speeds. | Lab | IC – TrueMax 2400 (Consentius Technologies, Sandy, UT, USA) | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: 20.33% |
| Koehler. 2011 | N=14 (0 F)<br>Age: 30.4 ± 6.2 y<br>BMI: 23.2 ± 1.4 kg/m$^2$ | 7-day comparison to DLW. | Field | DLW – 7 days | SenseWear Pro3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro3 Armband: -1.83% |
| Lee, 2011 | N=46 (21 F)<br>Age: 24.8 ± 5.6 y<br>BMI: 24.3 ± 3.6 kg/m$^2$ | Subjects completed 4-minute periods of standing, walking, jogging, and running. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | ePulse Personal Fitness Assistant (ePulse) (Impact Sports Technologies, San Diego, CA, USA) | Forearm | ePulse Personal Fitness Assistant -3.46% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lee, 2014 | N=60 (30 F)<br><br>Age: 26.4 y<br><br>BMI: 23.05 kg/m$^2$ | Subjects performed 13 activities for 5 minutes. Activities were categorized into sedentary, treadmill walking, treadmill jogging and moderate-to-vigorous activities (ascending and descending stairs, stationary bike, elliptical exercise, Wii tennis play, and basketball). | Lab | IC – Oxycon Mobile 5.0 (Erich Jaeger, Viasys Healthcare, Germany) | BodyMedia CORE (BodyMedia Inc., Pittsburgh, PA, USA)<br><br>Jawbone UP (Jawbone, San Francisco, California, USA)<br><br>Basis B1 Band (Basis Science Inc, San Francisco, CA, USA)<br><br>Nike Fuel Band (Nike Inc., Beaverton, OR, USA) | Upper arm<br><br>Wrist | BodyMedia CORE:-5.31%<br><br>Jawbone UP:<br><br>-6.92%<br><br>Basis B1 Band:<br><br>-31.65%<br><br>Nike Fuel Band: -1.91% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Lopez, 2017[1] | N=36 (16 F)<br>Age: 37.7 ± 9.8 y<br>BMI: 23.4 ± 2.8 kg/m$^2$ | Subjects performed a structured protocol including rest, computer use, standing, slow walking, running, basketball and overground cycling. | Lab | IC – MetaMax 3x (Cortex Biophysic, Leipzig, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Mini Armband: -16.00% |
| Mackey, 2011 | N=19 (8 F)<br>Age: 82 ± 3.3 y<br>BMI: 28.1 ± 3.8 kg/m$^2$ | 12.5-day comparison to DLW. | Field | DLW – 12.5 days | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: -0.05% |
| Martien, 2015 | N=60 (47 F)<br>Age: 85.5 ± 5.5 y<br>BMI: N/A | Subjects performed activity for 4 minutes and separated by 4 minutes seated rest. Activities included: Walking, rising and sitting in chairs positioned 5 meters apart and moving light objects. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Mini Armband: -12.00% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Maschac, 2013[1] | N=19 (13 F)<br>Age: 55.65 y<br>BMI: 31.5 ± 3.6 kg/m$^2$ | Subjects performed three walking sessions on a treadmill with different combinations of speed and incline. | Lab | IC – $VO_{2000}$ aerograph (Medgraphics, Saint Paul, MN, USA) | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband: 50.69% |
| McMinn, 2013 | N=19 (6 F)<br>Age: 30 y<br>BMI: 23.6 kg/m$^2$ | Subjects completed 3 treadmill walking trials at self-selected slow, medium, and fast speeds. | Lab | IC – Ultima CPX (Medgraphics, Saint Paul, MN, USA) | Actigraph GT3X+ (Actigraph Inc, Pensacola, FL, USA) | Wrist | Actigraph GT3X+ : -8.84% |
| Melanson, 2009 | N=7 (3 F)<br>Age: 31.8 ± 7.2 y<br>BMI: 27.8 ± 7.9 kg/m$^2$ | Subjects performed individualised protocols, including bench stepping and stationary cycling. | Lab | MC – 22.8 hours | LifeChek Calorie Sensor (LifeChek, LLC, Pittsburgh, PA, USA) | Wrist | LifeChek calorie sensor -4.87% |
| Soric, 2011 | N=19 (11 F)<br>Age: 28 ± 6 y<br>BMI: 23 ± 3 kg/m$^2$ | Subjects performed in-line skating exercises on a circular track at a self-selected pace. | Field | IC – COSMED K4b2 (COSMED, Rome, Italy) | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband : -73.33% |
| Montoye, 2017 | N=32 (14 F)<br>Age: 23.7 y<br>BMI: 25.5 kg/m$^2$ | Subjects completed 14 exercises, 11 in the laboratory including walking, jogging and cycling ergometry and 3 track exercises included self-paced walking at both a leisure and brisk pace for 200 meters and | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | Fitbit Charge HR (Fitbit Inc, San Francisco, California, USA) | Upper arm | Fitbit Charge HR: 7.59% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | self-paced jogging for 400 meters. Each was 5 minutes in duration. | | | | | |
| Murakami, 2016 | N=19 (10 F)<br>Age: N/A<br>BMI: N/A | 1) 12.5-day comparison to DLW.<br>2) 24 hours in metabolic chamber where subjects where subjects were required to perform deskwork, watch television, housework, treadmill walking, and sleeping. | Lab/ Field | DLW – 12.5 days<br>MC – 24 hours | Withings Pulse O2 (Withings, Issy-les-Moulineaux, France)<br>Garmin vivofit (Garmin ltd, Olathe, Kansas, USA)<br>Fitbit Flex (Fitbit Inc, San Francisco, California, USA)<br>Misfit Shine (Misfit, San Francisco, California, USA)<br>Epson Pulsense (Epson, Suwa, Nagano Prefecture, Japan) | Wrist | Withings Pulse O2: -22.03%<br>Garmin vivofit: -20.55%<br>Fitbit Flex: -1.04%<br>Misfit Shine: -2.36%<br>Epson Pulsense: -4.28% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Nelson, 2016 | N=30 (15 F)<br>Age: 48.9 ± 19.4 y<br>BMI: 26.3 ± 5.2 kg/m$^2$ | Subjects performed a structured protocol consisting of sedentary, household, and ambulatory activities. | Lab | IC – COSMED K4b2 (COSMED, Rome, Italy) | Jawbone UP (Jawbone, San Francisco, California, USA)<br>Fitbit Flex (Fitbit Inc, San Francisco, California, USA) | Wrist | Jawbone UP: -2.12%<br>Fitbit Flex: 12.74% |
| Papazoglou, 2006 | N=29<br>Age: N/A<br>BMI: N/A | Subjects performed a resting protocol in a larger sample and 29 of the obese subjects participated in low intensity modes of exercise including cycle ergometry, stair | Lab | IC – SensorMedics Vmax 229 (SensorMedics Inc, Yorba Linda, CA, USA) | SenseWear Pro 2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Wrist | SenseWear Pro 2 Armband: 21.54% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | stepping and treadmill walking. | | | | | |
| Price, 2017 | N=14 (3 F)<br>Age: 23 y<br>BMI: 22.8 kg/m$^2$ | Subjects walked on a treadmill at increasing velocities. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics, East Sandy, UT, USA) | Jawbone UP (Jawbone, San Francisco, California, USA)<br>Garmin vivofit (Garmin ltd, Olathe, Kansas, USA) | Upper arm | Jawbone UP: 56.91%<br>Garmin vivofit: 18.16% |
| Reece, 2015 | N=22 (11 F)<br>Age: N/A<br>BMI: N/A | Subjects performed a protocol including rest, sedentary activities and walking. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Wrist | SenseWear Mini Armband:<br>-3.79% |
| Reeve, 2014[1] | N: 18 (7 F)<br>Age: 22.6 y<br>BMI: 22.9 kg/m$^2$ | Subjects performed 2 resistance training sessions that included 9 different exercises. The weight lifted was 70% of 1 repetition max with 90-second rest intervals. | Lab | IC – COSMED K4b2 (COSMED, Rome, Italy) | BodyMedia CORE (BodyMedia Inc., Pittsburgh, PA, USA)<br>SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | BodyMedia CORE: 13.8%<br>SenseWear Mini Armband: 23.7% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Rousset, 2015 | Free-living: N=41 (20 F)<br>Lab: N=49 (26 F)<br>Age: N/A<br>BMI: N/A | 1) 10-day comparison to DLW.<br>2) 24 hours in metabolic chamber, which included eating, deskwork, watching television, housework, treadmill walking, and sleeping. | Lab/ Field | DLW – 12.5 days<br>MC – 17 hours | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband: -2.80% |
| Ryan, 2013 | N=26 (15 F)<br>Age: 24.7 y<br>BMI: 22.8 kg/m$^2$ | Subjects performed ambulatory activities on a treadmill. | Lab | IC – COSMED Quark CPNET (COSMED, Rome, Italy | SenseWear Pro 2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 2 Armband: -<br>-16.62% |
| Shcherbina, 2017[1] | N=60 (31 F)<br>Age: 38.5 y<br>BMI: 23.65 kg/m$^2$ | Subjects performed treadmill flat and incline running and cycle ergometry at low and moderate intensity. | Lab | IC – COSMED Quark CPNET (COSMED, Rome, Italy) | Apple watch (Apple Inc, Cupertino, CA, USA)<br>Basis Peak (Basis Science Inc, San Francisco, CA, USA) | Wrist | Apple watch: -38.23%<br>Basis Peak: -12.94%<br>Fitbit Surge: -3.86% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Fitbit surge (Fitbit Inc, San Francisco, CA, USA)<br><br>Microsoft band (Microsoft Corporation, Redmond, WA, USA)<br><br>PulseOn (PulseOn Oy, Espoo Finland) | | Microsoft Band<br>-19.64%<br>PulseOn: -24.47% |
| Slinde, 2013 | N=62 (62 F)<br>Age: 33.2 ± 4.2 y<br>BMI: 30 ± 2.8 kg/m$^2$ | 7-day comparison to DLW | Field | DLW – 7 days | SenseWear Pro 2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Wrist | SenseWear Pro 2 Armband: -2.90% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Smith, 2012 | N=30 (30 F)<br>Age: 29.0 ± 4.3 y<br>BMI: 24.1 ± 3.0 kg/m$^2$ | Subjects performed a series of activities of daily living activities and treadmill walking at increasing intensities. | Lab | IC – Parvo TrueOne 2400 (Parvo Medics East Sandy, UT, USA) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) Algorithm v2.2 | Upper arm | SenseWear Mini Armband: 18.43% |
| Stackpool, 2014 | N=20 (10 F)<br>Age: N/A<br>BMI: N/A | Subjects performed treadmill walking, treadmill running, elliptical exercise and an agility drills. | Lab | IC – Oxycon pro Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | Nike Fuel Band (Nike Inc, Beaverton, OR, USA)<br>Jawbone UP (Jawbone, San Francisco, California, USA)<br>Bodymedia Core (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | Nike Fuel Band: -3.99%<br>Jawbone UP: 3.09% |
| St-Onge, 2007 | N=45 (32 F)<br>Age: 35.1 ± 14 y<br>BMI: 23.9 ± 4.0 kg/m$^2$ | 10-day comparison to DLW. | Field | DLW – 10 days | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Armband: 4.70% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Tucker, 2015 | N=24 (13 F)<br>Age: 28.4 ± 7.8 y<br>BMI: 23.8 ± 3.9 kg/m$^2$ | Subjects performed two, 60-minute semi-structured routines consisting of sedentary/light-intensity, moderate-intensity and vigorous-intensity physical activity. | Lab | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | Nike Fuel Band (Nike Inc., Beaverton, OR, USA)<br>SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | Nike Fuel Band: 1.22%<br>SenseWear Armband: -2.10% |
| Van Helst, 2012 | N=21 (10 F)<br>Age: 29.3 ± 5.1 y | Subjects performed a treadmill protocol involving slow and moderate walking, running slowly, vigorously running and periods of rest. | Lab | IC – Gas analyzer (Respironics Novametrix Medical SystemW inc, NICO 7300, Wallingford, USA) | Vivago (Vivago Wellness, Paris, France) | Wrist | Vivago: -8.02% |
| Van Hoye, 2014 | N=44 (20 F)<br>Age: 21.1 ± 1.4 y<br>BMI: 21.8 ± 1.4 kg/m$^2$ | Subjects performed an incremental running test on a treadmill. | Lab | IC – Metalyzer 3B (Cortex Biophysic, Leipzig, Germany) | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband: -32.96% |
| Van Hoye, 2015 | N=39 (18 F) | Subjects performed exercise consisting of 5 minutes standing followed by alternating | Lab | IC – Metalyzer 3B (Cortex Biophysic, | SenseWear Armband (HealthWear, Bodymedia, | Upper arm | SenseWear Pro 3 Armband: - |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Age: 21.1 ± 1.4 y<br>BMI: 21.8 ± 1.4 kg/m$^2$ | walking and running at 35% and 65% $VO_{2max}$. | | Leipzig, Germany) | Pittsburgh, PA, USA)<br>Algorithm v2.2 | | -15.23% |
| | | | | | SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA)<br>Algorithm v5.2 | | |
| Vernillo, 2015 | N=20 (8 F)<br>Age: 30.1 ± 7.2 y<br>BMI: 22.1 ± 2.4 kg/m$^2$ | Subjects performed randomized pole walking activities at a constant speed and a variety of gradients. | Lab | IC – COSMED Quark b2 (COSMED, Rome, Italy) | SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm | SenseWear Pro 3 Armband:<br>-9.76% |
| | | | | | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | | SenseWear Mini Armband: -12.50 |
| Wahl, 2017 | N=20 (10 F)<br>Age: 25.2 y<br>BMI: 22.8 kg/m$^2$ | Subjects performed a running protocol consisting of four 5-minute stages of treadmill running at different velocities followed by a period of intermittent running | Lab/ Field | IC – Metalyzer 3B (Cortex Biophysic, Leipzig, Germany) | SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper arm/Wrist | SenseWear Mini Armband:<br>-21.27% |

and then a 2.4 km outdoor run.

Beurer AS80 (Beurer GmbH, Ulm, Germany)

Beurer AS80: -58.07%

Polar Loop (Polar Electro, Kempele, Finnland)

Polar Loop: 18.05%

Garmin vivofit (Garmin ltd, Olathe, Kansas, USA)

Garmin vivofit: -13.67%

Garmin vivosmart (Garmin ltd, Olathe, Kansas, USA)

Garmin vivosmart: 5.98%

Garmin vivoactive (Garmin ltd, Olathe, Kansas, USA)

Garmin vivoactive: 3.42%

Garmin Forerunner 920XT (Garmin ltd, Olathe, Kansas,

USA)

Fitbit Charge (Fitbit Inc, San Francisco, California, USA)

Garmin Forerunner 920XT:

-21.02%

Fitbit charge HR (Fitbit Inc, San Francisco, California, USA)

Fitbit Charge:

3.58%

Withings Pulse (Withings, Issy-les-Moulineaux, France)

Fitbit charge HR:

7.58%

Withings Pulse O2:

-15.98%

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Wallen 2016 | N=22 (11 F)<br>Age: 24.9 y<br>BMI: 24.3 kg/m$^2$ | Subjects performed a protocol including treadmill exercise and cycling ergometry. | Lab | IC – Metalyzer 3B (Cortex Biophysic, Leipzig, Germany) | Apple watch (Apple Inc, Cupertino, California, USA)<br><br>Fitbit charge HR (Fitbit Inc, San Francisco, California, USA)<br><br>Samsung Gear S (Samsung Electronics Co, Ltd, Suwon, South Korea)<br><br>Mio Alpha (Mio Global, | Wrist | Apple watch: -75.71<br><br>Fitbit charge HR:<br>-26.31%<br><br>Samsung Gear S:<br>-9.98%<br><br>Mio Alpha:<br>-53.19% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | Canada) | | |
| Woodman, 2017 | N=28 (8 F)<br>Age: 24.85 y<br>BMI: 24.25 kg/m$^2$ | Subjects performed a range of activities including: supine rest, household tasks, treadmill walking, stair stepping, outdoor walking, cycling, and running at a self-selected pace. Seated rest, and ergometer cycling. | Lab/ Field | IC – Oxycon Mobile portable metabolic system (Erich Jaeger, Viasys Healthcare, Germany) | Withings Pulse (Withings, Issy-les-Moulineaux, France)<br>Basis Peak (Basis Science Inc, San Francisco, CA, USA)<br>Garmin vivofit<br>(Garmin ltd, Olathe, Kansas, USA) | Wrist | Withings Pulse: -133.33%<br>Basis Peak: 0.59%<br>Garmin vivofit: -80.59% |

**Appendix 1.3 Device information**

| Device | Price | Wear site | Device grade | Input setup data | Sensors | Output | Battery life | Number of comparisons in meta-analysis | Weighted percent error |
|---|---|---|---|---|---|---|---|---|---|
| Actical (Phillips Respironics Inc, Murrysville, PN, USA) | €678 (incl. software)/ €321 (unit) | Hip, ankle, wrist | Research | Age, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Activity intensity<br>Kcals, steps | 194 days | 1 | |
| Actigraph GT3X+ (Actigraph Inc, Pensacola, FL, USA) | $250 | Hip, ankle, wrist | Research | Age, Gender, Race, H, W | Accelerometer: Triaxial<br>Heart rate: | Activity intensity<br>Kcals, sleep, steps | 31 days | 1 | -8.84% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Heat sensors: | | | | |
| Apple watch (Apple Inc, Cupertino, California, USA) | £249 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance tracking, Kcals, HR, minutes of brisk activity | 18 Hours | 4 | -6.59% |
| Apple watch 2 (Apple Inc, Cupertino, California, USA) | £315 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance tracking, Kcals, HR, minutes of brisk activity | 18 Hours | 1 | 48.20% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Basis b1 (Basis Science Inc, San Francisco, CA, USA) | £149 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: Yes | Steps, distance, Kcals, HR, active minutes, sleep | 5 days | 1 | -31.65% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Basis Peak (Basis Science Inc, San Francisco, CA, USA) | £170 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: Yes | Steps, distance, Kcals, HR, active minutes, sleep | 5 days | 1 | 0.59% |
| Beurer AS80 (Beurer GmbH, Ulm, Germany) | £29.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | 14 days | 1 | -58.07% |
| BodyMedia CORE (BodyMedia Inc., Pittsburgh, PA, | $150 | Upper left arm | Research (commercially | Age, Gender, H, W | Accelerometer: Triaxial | Steps, activity intensity, | 14 days | 4 | -1.06% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| USA) | | | available) | | | Kcals, sleep | | | |
| | | | | | Heart rate: | | | | |
| | | | | | Heat sensors: Yes | | | | |
| Epson Pulsense (Epson, Suwa, Nagano Prefecture, Japan) | £79.99 | Wrist | Commercial | Age, Gender, H, W, RHR | Accelerometer: Triaxial | Steps, distance, kcals, active minutes, HR, sleep | 36 hours | 1 | -4.28% |
| | | | | | Heart rate: Yes | | | | |
| | | | | | Heat sensors: | | | | |
| ePulse Personal Fitness Assistant (ePulse) (Impact Sports | $129.95 | Forearm | Commercial | Age, Gender, H, W, RHR | Accelerometer: Triaxial | Kcals, HR | | 1 | -3.46% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Technologies, San Diego, CA, USA) | | | | | Heart rate: Yes<br><br>Heat sensors: | | | | |
| Fitbit blaze (Fitbit Inc, San Francisco, California, USA | £134.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br><br>Heart rate: Yes<br><br>Heat sensors:<br><br>Triaxial accelerometer, altimeter, optical HR | Steps, distance, Kcals, active minutes, sleep, HR, steps | 5 days | 1 | 28.66% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Fitbit charge (Fitbit Inc, San Francisco, California, USA) | £109.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors:<br>Triaxial accelerometer, altimeter | Steps, distance, Kcals, active minutes, sleep | 5 days | 2 | -5.06% |
| Fitbit charge 2 (Fitbit Inc, San Francisco, California, USA) | £109.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep, HR, steps | 5 days | 1 | -30.97% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Fitbit charge HR (Fitbit Inc, San Francisco, California, USA) | £139.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep, HR, steps | 5 days | 6 | 1.3% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Fitbit Flex (Fitbit Inc, San Francisco, California, USA) | £79.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | 5 days | 5 | 8.22% |
| Fitbit Surge (Fitbit Inc, San Francisco, California, USA) | £289.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, active minutes, altimeter, GPS | 5 days | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Garmin Forerunner 225 (Garmin ltd, Olathe, Kansas, USA) | £199.99 | Wrist | Commercial | Age, Gender, H, W, RHR, HRmax | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, HR, distance, Kcals, active minutes, altimeter, GPS | 7-10 Hours | 1 | 44.23% |
| Garmin Forerunner 920XT (Garmin ltd, Olathe, Kansas, USA) | £450 | Wrist | Commercial | Age, Gender, H, W, RHR, HRmax | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, altimeter, sleep, HR, GPS | 3 Days | 1 | -21.02% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Garmin vivoactive (Garmin ltd, Olathe, Kansas, USA) | £250 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, altimeter, sleep, GPS | 7 Days | 1 | 3.42% |
| Garmin vivofit (Garmin ltd, Olathe, Kansas, USA) | £79.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | 1 Year | 5 | -26.09% |
| Garmin Vivosmart (Garmin ltd, Olathe, Kansas, USA) | £139.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: | Steps, distance, Kcals, active minutes, sleep | 7 Days | 1 | 5.98% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Heat sensors: | | | | |
| Garmin Vivosmart HR (Garmin ltd, Olathe, Kansas, USA) | £129.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, HR, intensity minutes, sleep | 7 Days | 1 | 16.85% |
| Jawbone UP (Jawbone, San Francisco, CA, USA) | £99.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: | Distance (app), Kcals, Steps, sleep | 10 days | 4 | 10.90% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Heat sensors: | | | | |
| Jawbone UP24 (Jawbone, San Francisco, CA, USA) | £89.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Distance (app), Kcal, Steps, sleep | 14 Days | 3 | -29.58% |
| LifeChek calorie sensor (LifeChek, LLC, Pittsburgh, PA, USA) | | Upper right arm | Commercial | | Accelerometer: Triaxial<br>Heart rate: | Kcals | | 1 | -4.87% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Heat sensors: Yes | | | | |
| Microsoft band (Microsoft Corporation, Redmond, WA, USA) | £169.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: Yes | Steps, distance, kcals, active minutes, sleep, HR, GPS | 48 Hours | 1 | -49.15% |
| Mio Alpha (Mio Global, Canada) | £119.99 | Wrist | Commercial | Age, Gender H, W, HRMAX, RHR | Accelerometer:<br>Heart rate: Yes<br>Heat sensors: | Kcals, HR | 24 Hours | 1 | -53.19% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Misfit Shine (Misfit, San Francisco, California, USA) | £99.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | | 1 | -2.36% |
| Nike Fuel Band (Nike Inc, Beaverton, OR, USA) | £129.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | 4 days | 3 | -0.48% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Polar Loop (Polar Electro, Kempele, Finnland) | £49.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep | 12 days | 1 | 18.05% |
| Polar: AW200 (Polar Electro Oy, Kempele, Finland | €152 (watch+softwar e) | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, distance, Kcals, active minutes | | 1 | -13.17% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Triaxial accelerometer | | | | |
| Polar: AW360 (Polar Electro Oy, Kempele, Finland | £149.99 | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep, HR | 12 Days | 1 | 28.68% |
| Samsung Gear S (Samsung Electronics Co, Ltd, Suwon, South Korea) | | Wrist | Commercial | Age, Gender, H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, Kcals, active minutes, sleep, HR, GPS | 2 Days | 1 | -9.98% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SenseWear Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | €800 (device)+ €1597 (software) | Upper right arm | Research | Age, Gender H, W, | Accelerometer: Biaxial<br>Heart rate:<br>Heat sensors: Yes | Steps, activity intensity, Kcals, sleep | 14 days | 12 | -4.31% |
| SenseWear Mini Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | | Upper left arm | Research | Age, Gender H, W, smoking status | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: Yes | Steps, activity intensity, Kcals, sleep | 28 days | 9 | -1.44% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SenseWear Pro 2 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper right arm | Research | Age, Gender H, W, smoking status | Accelerometer: Biaxial<br>Heart rate:<br>Heat sensors: Yes | Steps, activity intensity, kcal, sleep | 14 days | 7 | -7.54% |
| SenseWear Pro 3 Armband (HealthWear, Bodymedia, Pittsburgh, PA, USA) | Upper right arm | Research | Age, Gender H, W, smoking status | Accelerometer: Biaxial<br>Heart rate:<br>Heat sensors: Yes | Steps, activity intensity, kcal, sleep | 14 days | 12 | -4.56% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| TomTom Touch (TomTom, Amsterdam, the Netherlands) | £129.99 | Wrist | Commercial | Age, Gender H, W | Accelerometer: Triaxial<br>Heart rate: Yes<br>Heat sensors: | Steps, distance, activity intensity, Kcal, sleep, HR, | 5 Days | 1 | 28.66% |
| Vivago (Vivago WellnessW, Paris, France). | | Wrist | Commercial | | Accelerometer: Triaxial<br>Heart rate:<br>Heat sensors: | Steps, activity intensity, Kcal, sleep | | 1 | -8.02% |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Withings Pulse (Withings, Issy-les-Moulineaux, France) | £39.99 | Wrist, pocket or clip on | Commercial | Age, Gender H, | Accelerometer: Triaxial<br>Heart rate: (non continuous)<br>Heat sensors: | Steps, distance, Kcal, sleep | 14 days | 1 | -133.33% |
| Withings Pulse 02 (Withings, Issy-les-Moulineaux, France) | £79.99 | Wrist | Commercial | Age, Gender H, W | Accelerometer: Triaxial<br>Heart rate: (non | Steps, distance, activity intensity, | 14 days | 2 | -19.42% |

| | | |
|---|---|---|
| | continuous) | Kcal, sleep, |
| | Heat sensors: | HR, blood oxygen |

Characteristics of devices included in the meta-analysis. Weighted percentage error represents the sum of percentage difference multiplied by the relative weight within each meta-analysis.

Abbreviations: *Height (H), Weight (W), Kilocalories (Kcal), Heart Rate (HR), Global Positioning System (GPS).*

# Appendix 1.4 Risk of bias

| | Reporting (/11) | External validity (/3) | Internal validity (/4) |
|---|---|---|---|
| **Alsubheen, 2016** | 10 | 0 | 4 |
| **Bai, 2017** | 9 | 0 | 4 |
| **Benito, 2012** | 8 | 0 | 4 |
| **Berntsen, 2010** | 9 | 0 | 4 |
| **Berntsen, 2012** | 9 | 2 | 4 |
| **Bhammar, 2016** | 11 | 0 | 4 |
| **Boudreaux, 2018** | 10 | 0 | 4 |
| **Brazeau, 2011** | 10 | 0 | 4 |
| **Brazeau, 2014** | 11 | 0 | 3 |
| **Brazeau, 2016** | 11 | 1 | 4 |
| **Brugniaux, 2010** | 8 | 1 | 3 |
| **Calabro, 2014** | 9 | 0 | 4 |
| **Calabro, 2015** | 11 | 1 | 4 |
| **Casiraghi, 2013** | 11 | 0 | 4 |
| **Choudhry, 2017** | 9 | 0 | 4 |
| **Colbert, 2011** | 10 | 1 | 3 |
| **Correa, 2016** | 10 | 0 | 3 |
| **Diaz, 2015** | 7 | 0 | 4 |
| **Diaz, 2016** | 9 | 0 | 4 |
| **Dondzilla, 2016** | 8 | 0 | 4 |
| **Dooley, 2017** | 10 | 0 | 4 |
| **Drenowatz, 2011** | 9 | 0 | 4 |
| **Erdogan, 2010** | 9 | 0 | 3 |
| **Fruin, 2010** | 9 | 0 | 3 |
| **Furlanetto, 2010** | 11 | 0 | 4 |
| **Gastin, 2017** | 8 | 0 | 4 |

| | | | |
|---|---|---|---|
| **Heiermann, 2011** | 8 | 2 | 4 |
| **Imboden, 2017** | 9 | 0 | 4 |
| **Jakicic, 2004** | 10 | 0 | 4 |
| **Johannsen, 2010** | 9 | 1 | 4 |
| **Kim, 2015** | 8 | 0 | 4 |
| **King, 2004** | 9 | 0 | 4 |
| **Koehler, 2011** | 10 | 1 | 4 |
| **Lee, 2011** | 9 | 0 | 4 |
| **Lee, 2014** | 9 | 0 | 4 |
| **Mackey, 2011** | 11 | 3 | 4 |
| **Martien, 2015** | 9 | 2 | 4 |
| **McMinn, 2013** | 9 | 0 | 4 |
| **Melanson, 2009** | 5 | 0 | 2 |
| **Mikulic, 2011** | 10 | 0 | 4 |
| **Montoye, 2017** | 10 | 0 | 4 |
| **Murakami, 2016** | 7 | 1 | 4 |
| **Nelson, 2016** | 10 | 0 | 4 |
| **Papazoglou, 2006** | 9 | 0 | 4 |
| **Price, 2017** | 9 | 0 | 4 |
| **Reece, 2015** | 9 | 0 | 4 |
| **Rousset, 2015** | 9 | 1 | 4 |
| **Ryan, 2013** | 10 | 0 | 2 |
| **Slinde, 2013** | 10 | 2 | 4 |
| **Smith, 2012** | 10 | 0 | 4 |
| **St-Onge, 2007** | 9 | 1 | 3 |
| **Stackpool, 2015** | 9 | 0 | 4 |
| **Tucker, 2015** | 11 | 0 | 4 |
| **Van helst, 2012** | 9 | 0 | 4 |
| **Van Hoye, 2014** | 9 | 0 | 4 |

| | | | |
|---|---|---|---|
| **Van Hoye, 2015** | 10 | 0 | 4 |
| **Vernillo, 2015** | 8 | 0 | 4 |
| **Wahl, 2017** | 9 | 0 | 4 |
| **Wallen 2016** | 9 | 0 | 4 |
| **Woodman, 2017** | 8 | 0 | 4 |

# Appendix 2.1 Simulation study results

| | | NoHoW | | | | | | | Multiple Imputation | | | | | | | Kalman | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | **Window** | **Mean** | **SD** | **Difference** | **RMSE mean** | **RMSE median** | **RMSE minimum** | **RMSE maximum** | **Mean** | **SD** | **Difference** | **RMSE mean** | **RMSE median** | **RMSE minimum** | **RMSE maximum** | **Mean** | **SD** | **Difference** | **RMSE mean** | **RMSE median** | **RMSE minimum** | **RMSE maximum** |
| **TDEE** | **1** | 2653.42 | 0.95 | 26.83 | 31.14 | 30.94 | 28.82 | 33.12 | 2642.21 | 1.01 | 15.63 | 21.30 | 21.08 | 19.20 | 23.11 | 2658.34 | 0.95 | 31.75 | 37.44 | 37.29 | 35.49 | 39.90 |
| | **2** | 2653.27 | 1.74 | 26.69 | 33.37 | 33.54 | 30.89 | 36.53 | 2639.92 | 1.80 | 13.33 | 24.63 | 24.42 | 21.95 | 28.31 | 2658.06 | 1.53 | 31.47 | 39.30 | 39.43 | 35.24 | 43.27 |
| | **3** | 2652.90 | 2.19 | 26.31 | 36.52 | 36.55 | 31.54 | 40.64 | 2637.42 | 2.63 | 10.84 | 28.65 | 28.47 | 23.87 | 34.02 | 2658.35 | 2.43 | 31.76 | 42.20 | 42.46 | 34.15 | 45.32 |
| | **4** | 2653.45 | 2.50 | 26.86 | 39.01 | 39.11 | 33.44 | 43.29 | 2634.83 | 2.64 | 8.25 | 32.33 | 32.23 | 28.66 | 37.27 | 2658.98 | 2.10 | 32.39 | 46.42 | 46.05 | 43.11 | 53.72 |
| | **5** | 2653.59 | 2.65 | 27.00 | 43.30 | 42.86 | 39.59 | 50.00 | 2633.03 | 3.02 | 6.44 | 37.52 | 37.34 | 30.65 | 41.80 | 2657.87 | 2.95 | 31.28 | 48.60 | 49.33 | 41.11 | 54.41 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **6** | 2653.22 | 3.94 | 26.64 | 45.30 | 45.26 | 37.85 | 54.86 | 2630.24 | 4.50 | 3.65 | 39.75 | 40.02 | 36.01 | 44.82 | 2657.30 | 3.59 | 30.71 | 48.57 | 48.94 | 42.93 | 53.31 |
| | **7** | 2652.13 | 3.60 | 25.55 | 48.05 | 48.59 | 43.40 | 53.25 | 2628.40 | 3.37 | 1.81 | 46.10 | 45.96 | 41.59 | 51.91 | 2658.16 | 3.39 | 31.58 | 53.54 | 53.27 | 47.40 | 58.81 |
| | **8** | 2650.41 | 2.95 | 23.82 | 49.91 | 49.80 | 42.38 | 55.95 | 2622.22 | 3.73 | -4.37 | 48.78 | 48.99 | 43.34 | 57.54 | 2657.17 | 4.91 | 30.59 | 54.70 | 55.14 | 44.74 | 64.58 |
| | **9** | 2653.40 | 4.35 | 26.81 | 55.51 | 55.69 | 47.55 | 61.92 | 2622.73 | 4.41 | -3.86 | 52.79 | 53.30 | 44.13 | 57.52 | 2658.04 | 5.00 | 31.45 | 59.39 | 59.85 | 51.06 | 70.90 |
| | **10** | 2651.04 | 4.04 | 24.46 | 59.07 | 59.05 | 51.20 | 68.89 | 2619.06 | 3.90 | -7.52 | 59.20 | 60.90 | 48.71 | 68.05 | 2658.21 | 4.34 | 31.62 | 64.60 | 65.15 | 54.88 | 72.55 |
| **Steps** | **1** | 1080.61 | 13.05 | 235.27 | 291.90 | 291.66 | 271.60 | 317.28 | 1069.98 | 13.10 | 124.64 | 206.34 | 205.75 | 187.75 | 232.66 | 1076.98 | 13.12 | 193.64 | 267.59 | 266.09 | 246.41 | 296.26 |
| | **2** | 1080.20 | 24.73 | 231.86 | 330.03 | 329.13 | 302.91 | 360.77 | 1066.89 | 22.33 | 90.55 | 260.79 | 256.41 | 224.08 | 338.79 | 1076.57 | 20.11 | 191.23 | 316.96 | 313.37 | 281.82 | 366.37 |
| | **3** | 1080.53 | 26.97 | 230.18 | 377.08 | 376.18 | 328.60 | 434.21 | 1062.34 | 28.23 | 56.00 | 311.35 | 312.92 | 262.80 | 376.98 | 1076.55 | 30.02 | 198.21 | 388.04 | 394.35 | 322.77 | 427.97 |
| | **4** | 1080.87 | 32.32 | 238.53 | 417.16 | 411.94 | 361.67 | 472.68 | 1059.14 | 34.71 | 19.80 | 372.87 | 358.84 | 308.90 | 443.73 | 1077.31 | 28.40 | 206.97 | 451.00 | 448.51 | 418.41 | 495.35 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **5** | 1081 | 30. | 241.26 | 474. | 472. | 420.7 | 527.34 | 1056 | 33. | -2.34 | 437. | 436. | 396.3 | 484.87 | 1076 | 34. | 196.48 | 498. | 497. | 443.0 | 571.46 |
| | | 1.60 | 96 | | 52 | 84 | 8 | | 8.00 | 75 | | 90 | 41 | 1 | | 6.82 | 30 | | 83 | 95 | 8 | |
| | **6** | 1079 | 45. | 229.43 | 500. | 501. | 433.7 | 550.42 | 1052 | 49. | -48.92 | 476. | 470. | 410.5 | 582.31 | 1076 | 45. | 198.54 | 523. | 522. | 475.8 | 604.44 |
| | | 9.77 | 28 | | 57 | 42 | 4 | | 1.42 | 32 | | 90 | 89 | 1 | | 8.88 | 63 | | 26 | 33 | 3 | |
| | **7** | 1080 | 52. | 231.39 | 550. | 550. | 494.6 | 606.38 | 1050 | 49. | -61.28 | 551. | 543. | 457.3 | 637.15 | 1077 | 44. | 204.42 | 595. | 588. | 511.0 | 681.06 |
| | | 1.73 | 73 | | 51 | 33 | 7 | | 9.06 | 57 | | 31 | 59 | 8 | | 4.76 | 13 | | 22 | 24 | 2 | |
| | **8** | 1077 | 48. | 204.38 | 560. | 554. | 508.3 | 663.75 | 1042 | 59. | - | 611. | 601. | 551.3 | 762.66 | 1076 | 66. | 195.22 | 602. | 599. | 520.6 | 749.52 |
| | | 4.72 | 87 | | 91 | 54 | 5 | | 7.40 | 47 | 142.94 | 23 | 73 | 5 | | 5.56 | 49 | | 33 | 08 | 7 | |
| | **9** | 1080 | 55. | 238.90 | 629. | 625. | 560.6 | 704.13 | 1042 | 49. | - | 649. | 638. | 587.0 | 727.25 | 1077 | 74. | 207.92 | 690. | 686. | 580.7 | 829.83 |
| | | 9.24 | 23 | | 84 | 40 | 1 | | 8.41 | 33 | 141.93 | 65 | 63 | 5 | | 8.26 | 12 | | 48 | 65 | 0 | |
| | **10** | 1078 | 67. | 217.35 | 692. | 700. | 590.0 | 797.25 | 1038 | 51. | - | 718. | 738. | 618.8 | 844.70 | 1077 | 80. | 206.26 | 771. | 755. | 672.6 | 930.92 |
| | | 7.69 | 94 | | 79 | 76 | 1 | | 7.58 | 53 | 182.76 | 92 | 28 | 2 | | 6.60 | 49 | | 70 | 71 | 5 | |
| **Seden** | **1** | 1105. | 0.2 | 18.05 | 19.4 | 19.4 | 19.04 | 19.83 | 1112. | 0.3 | 24.29 | 25.2 | 25.1 | 24.51 | 25.90 | 1101. | 0.2 | 13.99 | 16.1 | 16.1 | 15.74 | 16.66 |
| **tary** | | 81 | 4 | | 0 | 0 | | | 04 | 5 | | 4 | 9 | | | 75 | 7 | | 5 | 6 | | |
| | **2** | 1105. | 0.4 | 18.08 | 19.6 | 19.6 | 18.74 | 20.50 | 1115. | 0.5 | 27.29 | 28.5 | 28.4 | 27.52 | 29.37 | 1101. | 0.4 | 14.18 | 16.9 | 16.8 | 16.09 | 17.56 |
| | | 84 | 5 | | 6 | 8 | | | 04 | 1 | | 0 | 0 | | | 93 | 1 | | 2 | 9 | | |
| | **3** | 1105. | 0.4 | 18.16 | 20.0 | 20.1 | 19.15 | 20.88 | 1118. | 0.7 | 30.40 | 31.8 | 31.8 | 30.54 | 33.01 | 1102. | 0.6 | 14.27 | 17.6 | 17.8 | 16.73 | 18.41 |
| | | 92 | 1 | | 8 | 5 | | | 16 | 2 | | 7 | 6 | | | 02 | 9 | | 9 | 1 | | |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **4** | 1105.97 | 0.66 | 18.21 | 20.43 | 20.43 | 19.55 | 21.28 | 1121.32 | 0.84 | 33.57 | 35.49 | 35.50 | 33.48 | 36.95 | 1101.84 | 0.72 | 14.08 | 18.10 | 18.05 | 16.94 | 19.54 |
| | **5** | 1105.87 | 0.67 | 18.11 | 20.97 | 20.97 | 19.80 | 21.93 | 1124.17 | 0.85 | 36.41 | 38.86 | 38.55 | 37.60 | 41.09 | 1102.40 | 0.98 | 14.64 | 19.32 | 19.34 | 16.97 | 20.84 |
| | **6** | 1105.96 | 0.85 | 18.20 | 21.44 | 21.40 | 20.13 | 22.67 | 1127.17 | 1.23 | 39.42 | 42.06 | 42.27 | 39.48 | 43.82 | 1102.69 | 1.09 | 14.93 | 20.21 | 20.34 | 18.87 | 21.55 |
| | **7** | 1106.10 | 0.76 | 18.34 | 22.04 | 22.00 | 20.54 | 23.70 | 1129.74 | 0.96 | 41.98 | 45.19 | 45.38 | 42.45 | 47.14 | 1102.47 | 1.27 | 14.71 | 20.85 | 20.90 | 18.91 | 22.17 |
| | **8** | 1106.79 | 1.02 | 19.03 | 23.19 | 23.09 | 21.38 | 25.75 | 1133.91 | 1.18 | 46.15 | 49.67 | 49.45 | 48.27 | 51.61 | 1102.66 | 1.13 | 14.90 | 21.69 | 21.80 | 18.68 | 23.86 |
| | **9** | 1105.58 | 1.13 | 17.83 | 22.94 | 23.10 | 21.07 | 24.62 | 1135.52 | 1.43 | 47.77 | 51.68 | 51.62 | 48.97 | 53.99 | 1103.27 | 1.55 | 15.52 | 22.92 | 23.00 | 20.12 | 24.79 |
| | **10** | 1106.56 | 1.03 | 18.80 | 24.87 | 24.89 | 23.15 | 26.39 | 1138.91 | 1.01 | 51.15 | 55.56 | 55.56 | 53.69 | 57.76 | 1102.98 | 1.39 | 15.22 | 23.73 | 23.84 | 21.46 | 26.89 |
| **Light** | **1** | 274.67 | 0.19 | 7.91 | 9.30 | 9.34 | 8.73 | 9.60 | 270.29 | 0.28 | 3.52 | 5.80 | 5.79 | 5.30 | 6.24 | 278.66 | 0.26 | 11.90 | 13.54 | 13.55 | 13.14 | 14.02 |
| | **2** | 274.58 | 0.38 | 7.81 | 9.60 | 9.64 | 8.72 | 10.23 | 268.08 | 0.42 | 1.31 | 6.20 | 6.29 | 5.04 | 7.36 | 279.13 | 0.48 | 12.37 | 14.62 | 14.60 | 13.82 | 16.02 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **3** | 274.60 | 0.34 | 7.83 | 10.11 | 10.03 | 9.51 | 10.82 | 265.93 | 0.60 | -0.84 | 7.70 | 7.73 | 6.63 | 8.52 | 279.76 | 0.62 | 13.00 | 15.96 | 15.97 | 14.90 | 16.98 |
| | **4** | 274.46 | 0.46 | 7.69 | 10.44 | 10.32 | 9.53 | 11.39 | 263.64 | 0.66 | -3.13 | 9.93 | 9.96 | 8.18 | 11.05 | 280.56 | 0.66 | 13.79 | 17.13 | 17.26 | 15.70 | 18.51 |
| | **5** | 274.57 | 0.63 | 7.80 | 11.19 | 11.06 | 10.12 | 12.36 | 261.60 | 0.68 | -5.17 | 12.45 | 12.43 | 10.77 | 14.50 | 280.68 | 0.93 | 13.91 | 17.82 | 17.91 | 15.65 | 19.86 |
| | **6** | 274.49 | 0.69 | 7.72 | 11.73 | 11.63 | 10.47 | 13.36 | 259.57 | 1.01 | -7.20 | 14.53 | 14.54 | 12.46 | 15.95 | 281.23 | 1.03 | 14.46 | 19.11 | 18.90 | 17.10 | 21.23 |
| | **7** | 274.52 | 0.55 | 7.75 | 12.39 | 12.37 | 11.29 | 13.60 | 257.82 | 0.72 | -8.95 | 16.90 | 16.89 | 14.92 | 19.96 | 282.05 | 1.24 | 15.29 | 20.70 | 20.78 | 18.32 | 22.27 |
| | **8** | 273.95 | 1.00 | 7.18 | 12.74 | 12.70 | 11.55 | 14.18 | 254.63 | 1.05 | -12.14 | 20.11 | 19.81 | 18.33 | 21.90 | 282.51 | 1.19 | 15.74 | 21.63 | 21.67 | 20.01 | 23.61 |
| | **9** | 274.84 | 1.08 | 8.08 | 14.18 | 14.24 | 11.86 | 16.32 | 253.69 | 1.23 | -13.07 | 21.73 | 22.01 | 18.96 | 23.77 | 282.56 | 1.50 | 15.80 | 22.50 | 22.68 | 19.90 | 24.15 |
| | **10** | 274.35 | 0.93 | 7.58 | 15.19 | 15.11 | 12.81 | 17.42 | 251.46 | 0.99 | -15.31 | 24.51 | 24.48 | 21.41 | 26.54 | 283.66 | 1.25 | 16.89 | 24.41 | 24.32 | 21.89 | 26.33 |
| **Moderate** | **1** | 51.96 | 0.09 | 1.72 | 2.33 | 2.34 | 2.04 | 2.53 | 50.41 | 0.10 | 0.16 | 1.16 | 1.16 | 0.95 | 1.42 | 51.98 | 0.10 | 1.74 | 2.89 | 2.88 | 2.59 | 3.15 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 52.01 | 0.17 | 1.77 | 2.73 | 2.72 | 2.43 | 3.04 | 49.75 | 0.16 | -0.49 | 1.94 | 1.90 | 1.50 | 2.54 | 51.41 | 0.17 | 1.17 | 2.90 | 2.96 | 2.19 | 3.39 |
| **3** | 51.95 | 0.22 | 1.71 | 3.01 | 3.01 | 2.26 | 3.56 | 48.94 | 0.24 | -1.30 | 2.87 | 2.91 | 2.31 | 3.51 | 50.83 | 0.22 | 0.59 | 3.05 | 3.00 | 2.53 | 4.06 |
| **4** | 51.99 | 0.29 | 1.75 | 3.24 | 3.24 | 2.59 | 3.78 | 48.27 | 0.25 | -1.97 | 3.56 | 3.59 | 2.95 | 4.23 | 50.32 | 0.29 | 0.08 | 3.50 | 3.53 | 3.08 | 4.06 |
| **5** | 51.96 | 0.27 | 1.72 | 3.70 | 3.68 | 3.18 | 4.69 | 47.60 | 0.28 | -2.64 | 4.54 | 4.60 | 4.10 | 4.93 | 49.73 | 0.36 | -0.51 | 3.99 | 3.96 | 3.31 | 4.57 |
| **6** | 51.97 | 0.31 | 1.73 | 4.05 | 4.03 | 3.49 | 4.94 | 46.80 | 0.40 | -3.44 | 5.51 | 5.63 | 4.60 | 6.21 | 49.06 | 0.36 | -1.18 | 4.50 | 4.40 | 3.94 | 5.30 |
| **7** | 51.90 | 0.40 | 1.66 | 4.19 | 4.20 | 3.72 | 4.92 | 46.14 | 0.35 | -4.11 | 6.34 | 6.29 | 5.47 | 6.92 | 48.57 | 0.44 | -1.67 | 5.01 | 5.01 | 4.60 | 5.68 |
| **8** | 51.71 | 0.40 | 1.47 | 4.56 | 4.52 | 3.97 | 5.48 | 45.31 | 0.40 | -4.94 | 7.47 | 7.35 | 6.53 | 8.36 | 47.96 | 0.43 | -2.28 | 5.74 | 5.62 | 4.89 | 7.20 |
| **9** | 51.98 | 0.37 | 1.74 | 5.04 | 5.04 | 4.28 | 5.68 | 44.76 | 0.38 | -5.48 | 8.03 | 8.05 | 7.21 | 8.75 | 47.44 | 0.54 | -2.80 | 6.14 | 6.16 | 5.33 | 7.07 |
| **10** | 51.63 | 0.46 | 1.39 | 5.38 | 5.23 | 4.72 | 6.26 | 43.85 | 0.33 | -6.39 | 9.14 | 9.06 | 8.68 | 10.10 | 46.80 | 0.50 | -3.44 | 7.25 | 7.28 | 6.40 | 8.17 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Vigorous** | **1** | 7.56 | 0.04 | 0.27 | 0.63 | 0.63 | 0.53 | 0.78 | 7.26 | 0.03 | -0.03 | 0.39 | 0.35 | 0.26 | 0.65 | 7.61 | 0.03 | 0.32 | 0.81 | 0.80 | 0.71 | 1.01 |
| | **2** | 7.57 | 0.07 | 0.28 | 0.83 | 0.83 | 0.69 | 0.99 | 7.12 | 0.07 | -0.17 | 0.70 | 0.62 | 0.48 | 1.13 | 7.52 | 0.08 | 0.23 | 0.97 | 0.94 | 0.78 | 1.31 |
| | **3** | 7.54 | 0.08 | 0.25 | 1.01 | 1.03 | 0.74 | 1.20 | 6.97 | 0.09 | -0.32 | 1.01 | 1.02 | 0.74 | 1.33 | 7.39 | 0.09 | 0.09 | 1.01 | 1.01 | 0.85 | 1.23 |
| | **4** | 7.58 | 0.10 | 0.29 | 1.25 | 1.25 | 0.94 | 1.48 | 6.77 | 0.07 | -0.52 | 1.36 | 1.33 | 1.12 | 1.93 | 7.28 | 0.08 | -0.01 | 1.32 | 1.31 | 1.02 | 1.76 |
| | **5** | 7.60 | 0.11 | 0.31 | 1.44 | 1.40 | 1.08 | 1.80 | 6.64 | 0.09 | -0.65 | 1.50 | 1.44 | 1.27 | 1.95 | 7.19 | 0.12 | -0.10 | 1.40 | 1.41 | 1.09 | 1.74 |
| | **6** | 7.58 | 0.14 | 0.29 | 1.58 | 1.48 | 1.40 | 1.94 | 6.46 | 0.12 | -0.83 | 1.85 | 1.82 | 1.44 | 2.18 | 7.03 | 0.13 | -0.26 | 1.58 | 1.54 | 1.26 | 2.13 |
| | **7** | 7.48 | 0.16 | 0.19 | 1.64 | 1.66 | 1.24 | 1.99 | 6.31 | 0.17 | -0.98 | 2.13 | 2.15 | 1.68 | 2.62 | 6.91 | 0.16 | -0.38 | 1.88 | 1.86 | 1.36 | 2.73 |
| | **8** | 7.55 | 0.13 | 0.26 | 1.81 | 1.82 | 1.44 | 2.13 | 6.16 | 0.14 | -1.13 | 2.33 | 2.32 | 1.64 | 2.85 | 6.87 | 0.15 | -0.42 | 1.92 | 1.90 | 1.41 | 2.38 |
| | **9** | 7.60 | 0.24 | 0.31 | 2.05 | 2.06 | 1.37 | 2.29 | 6.02 | 0.19 | -1.27 | 2.59 | 2.59 | 1.90 | 3.04 | 6.73 | 0.16 | -0.56 | 2.21 | 2.15 | 1.57 | 3.29 |
| | **10** | 7.47 | 0.3 | 0.18 | 2.25 | 2.24 | 1.84 | 3.03 | 5.79 | 0.2 | -1.50 | 2.91 | 2.82 | 2.38 | 3.71 | 6.57 | 0.1 | -0.72 | 2.28 | 2.21 | 1.85 | 2.95 |

| 1 | 3 | 5 |
| --- | --- | --- |

# Appendix 3.1 Algorithm hyperparameters

**Algorithm parameters**

**Fitbit Regression models**

**Study 1 as training**

***Random Forest***

*{'n_estimators': 600,*
*'min_samples_split': 2,*
*'min_samples_leaf': 4,*
*'max_features': 'sqrt',*
*'max_depth': 60,*
*'bootstrap': True}*

***Gradient boost***

*{'n_estimators': 1000,*
*'min_samples_split': 6,*
*'min_samples_leaf': 2,*
*'max_features': 'log2',*
*'max_depth': 20,*
*'loss': 'ls',*
*'learning_rate': 0.01}*

***Neural network***

*{'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

**Study 2 as training**

**Random forest**

*{'n_estimators': 200,*
*'min_samples_split': 2,*
*'min_samples_leaf': 2,*
*'max_features': 'sqrt',*
*'max_depth': 30,*
*'bootstrap': True}*

**Gradient boost**

*{'n_estimators': 1000,*
*'min_samples_split': 4,*
*'min_samples_leaf': 8,*
*'max_features': 'auto',*
*'max_depth': 2,*
*'loss': 'ls',*
*'learning_rate': 0.01}*

**Neural network**

*{'n_neurons': 15,*
*'n_hidden': 3,*
*'learning_rate': 0.0001}*

***LOSO***

*Random forest*

*{'n_estimators': 800,*
*'min_samples_split': 10,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 60,*

*'bootstrap': True}*

***Gradient boost***

*{'n_estimators': 1000,*
*'min_samples_split': 10,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 15,*

*'loss': 'ls',*

*'learning_rate': 0.01}*

***Neural network***

*{'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

**Sensewear regression models**

**Study 1 as training**

Random forest

*{ 'n_estimators': 200,*
*'min_samples_split': 5,*

*'min_samples_leaf': 4,*

*'max_features': 'sqrt',*

*'max_depth': 40,*

*'bootstrap': True}*

**Gradient boost**

*{ 'n_estimators': 10000,*
*'min_samples_split': 10,*

*'min_samples_leaf': 1,*

*'max_features': 'log2',*

*'max_depth': 15,*

*'learning_rate': 0.01}*

**Neural network**

*{'n_neurons': 15,*
*'n_hidden': 3,*
*'learning_rate': 0.0001}*

**Study 2 as training**

**Random forest**

*{'n_estimators': 800,*
*'min_samples_split': 10,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 60,*

*'bootstrap': True}*

**Gradient boost**

*{'n_estimators': 1000,*
*'min_samples_split': 10,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 15,*

*'learning_rate': 0.01}*

**Neural network**

*{'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

**LOSO**

***Random forest***

*{'n_estimators': 800,*

*'min_samples_split': 5,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 30,*

*'bootstrap': True}*

***Gradient boost***

*{'n_estimators': 10000,*

*'min_samples_split': 6,*

*'min_samples_leaf': 12,*

*'max_features': 'log2',*

*'max_depth': 15,*

*'learning_rate': 0.01}*

***Neural network***

*{'n_neurons': 15,*
*'n_hidden': 3,*
*'learning_rate': 0.0001}*

**ActiGraph Regression models**

**Study 1 as training**

***Random Forest***

*{'n_estimators': 1000,*
*'min_samples_split': 5,*

*'min_samples_leaf': 2,*

*'max_features': 'sqrt',*

*'max_depth': 50,*

*'bootstrap': False}*

***Gradient Boost***

*{'n_estimators': 10000,*
*'min_samples_split': 10,*

*'min_samples_leaf': 12,*

*'max_features': 'sqrt',*

*'max_depth': 10,*

*'loss': 'ls',*

*'learning_rate': 0.01}*

***Neural Network***

*{'n_neurons': 55,*

*'n_hidden': 3,*

*'learning_rate': 0.003}*

***Study 2 as training***

*Random forest*

*{'n_estimators': 200,*
*'min_samples_split': 2,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 60,*

*'bootstrap': True}*

***Gradient boosting***

*{'n_estimators': 10000,*
*'min_samples_split': 20,*
*'min_samples_leaf': 2,*
*'max_features': 'sqrt',*
*'max_depth': 20,*
*'learning_rate': 0.01}*

***Neural network***

*{ 'n_neurons': 65, 'n_hidden': 2, 'learning_rate': 0.0001}*

**LOSO**

*Random forest*

*{'n_estimators': 1000,*
*'min_samples_split': 5,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 60,*

*'bootstrap': False}*

***Gradient Boost***

*{'n_estimators': 10000,*

*'min_samples_split': 6,*

*'min_samples_leaf': 12,*

*'max_features': 'log2',*

*'max_depth': 15,*

*'learning_rate': 0.01}*

***Neural Network***

*{ 'n_neurons': 55,*
*'n_hidden': 2,*
*'learning_rate': 0.003*
*}*

**Classification models**

**Fitbit classification models**

**Study 1 as training**

***KNN***

*{'weights': 'uniform',*

*'n_neighbors': 15,*

*'metric': 'minkowski'}*

***Random forest***

*{'n_estimators': 1000,*
*'min_samples_split': 10,*
*'min_samples_leaf': 4,*
*'max_features': 'auto',*
*'max_depth': 80,*
*'bootstrap': True}*

***Gradient boost***

*{'n_estimators': 1000,*
*'min_samples_split': 2,*
*'min_samples_leaf': 1,*
*'max_features': 'log2',*
*'max_depth': 15,*
*'learning_rate': 0.1}*

***Neural network***

*{ 'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

***SVM***

*{ 'C': 5.046137691733707,*
*'gamma': 0.19767211400638388}*

***Study 2 as training***

***KNN***

*{'weights': 'uniform',*

*'n_neighbors': 33,*

*'metric': 'manhattan'}*

***Random forest***

*{'n_estimators': 800,*
*'min_samples_split': 5,*
*'min_samples_leaf': 4,*
*'max_features': 'auto',*
*'max_depth': 80,*
*'bootstrap': True}*

***Gradient boost***

*{'n_estimators': 1000,*
*'min_samples_split': 6,*
*'min_samples_leaf': 2,*
*'max_features': 'sqrt',*
*'max_depth': 2,*
*'learning_rate': 0.01}*

***Neural network***

*{ 'n_neurons': 15, 'n_hidden': 3, 'learning_rate': 0.0001}*

***SVM***

*{ 'C': 7.924145688620425, 'gamma': 0.14645041271999773}*

**LOSO**

***KNN***

*{ 'weights': 'distance', 'n_neighbors': 73, 'metric': 'manhattan'}*

***Random forest***

*{ 'n_estimators': 400,*
*'min_samples_split': 5,*
*'min_samples_leaf': 4,*
*'max_features': 'sqrt',*
*'max_depth': 70,*
*'bootstrap': True}*

***Gradient boost***

*{ 'n_estimators': 1000,*
*'min_samples_split': 4,*
*'min_samples_leaf': 8,*
*'max_features': 'auto',*
*'max_depth': 2,*
*'learning_rate': 0.01}*

***Neural network***

*{ 'n_neurons': 70, 'n_hidden': 2, 'learning_rate': 0.1}*

***SVM***

*{'C': 6.319450186421157, 'gamma': 0.3912291401980419}*

**Sensewear classification models**

**Study 1 as training**

**KNN**

{*'weights': 'uniform',*
*'n_neighbors': 27,*
*'metric': 'manhattan'*}

**Random Forest**

{*'n_estimators': 800,*
*'min_samples_split': 10,*
*'min_samples_leaf': 4,*

*'max_features': 'auto',*
*'max_depth': 40,*
*'bootstrap': False*}

**Gradient boost**

{*'n_estimators': 1000,*
*'min_samples_split': 4,*
*'min_samples_leaf': 4,*
*'max_features': 'log2',*
*'max_depth': 10,*
*'learning_rate': 0.1*}

**Neural network**

*{ 'n_neurons': 15, 'n_hidden': 3, 'learning_rate': 0.0001}*

**SVM**

*{ 'C': 7.924145688620425, 'gamma': 0.14645041271999773}*

**Study 2 as training**

**KNN**

*{'weights': 'uniform', 'n_neighbors': 53, 'metric': 'manhattan'}*

**Random forest**

*{ 'n_estimators': 1000,*
*'min_samples_split': 5,*
*'min_samples_leaf': 4,*
*'max_features': 'sqrt',*
*'max_depth': 40,*
*'bootstrap': True}*

**Gradient boost**

{ 'n_estimators': 10000,
'min_samples_split': 6,
'min_samples_leaf': 12,
'max_features': 'sqrt',
'max_depth': 20,
'learning_rate': 0.1
}

**Neural network**

*{ 'n_neurons': 15, 'n_hidden': 3, 'learning_rate': 0.0001}*

**SVM**

*{ 'C': 7.924145688620425, 'gamma': 0.14645041271999773}*

**LOSO**

***KNN***

*{'weights': 'uniform',*
*'n_neighbors': 33,*
*'metric': 'manhattan'}*

***Random forest***

*'n_estimators': 400,*
*'min_samples_split': 5,*
*'min_samples_leaf': 4,*
*'max_features': 'sqrt',*

*'max_depth': 70,*

*'bootstrap': True*

*}*

**Gradient boost**

*{ 'n_estimators': 1000,*
*'min_samples_split': 4,*

*'min_samples_leaf': 2,*

*'max_features': 'log2',*

*'max_depth': 10,*

*'learning_rate': 0.05*

*}*

**Neural network**

*{'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001*
*}*

**SVM**

*{ 'C': 7.924145688620425,*
*'gamma': 0.14645041271999773}*

**ActiGraph classification models**
**Study 1 as training**

**KNN**

*{'weights': 'uniform',*

*'n_neighbors': 41,*

*'metric': 'manhattan'}*

**Random forest**

*{'n_estimators': 200,*

*'min_samples_split': 2,*

*'min_samples_leaf': 1,*

*'max_features': 'sqrt',*

*'max_depth': 90,*

*'bootstrap': False}*

**Gradient boost**

*{'n_estimators': 1000,*

*'min_samples_split': 20,*

*'min_samples_leaf': 4,*

*'max_features': 'sqrt',*

*'max_depth': 10,*

*'learning_rate': 0.1}*

**Neural network**

*{ 'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

**SVM**

{ 'C': 7.924145688620425,
'gamma': 0.14645041271999773}

**Study 2 as training**

**KNN**

{ 'weights': 'uniform', 'n_neighbors': 17, 'metric': 'manhattan'}

**Random forest**

{ 'n_estimators': 800,
'min_samples_split': 5,

'min_samples_leaf': 4,

'max_features': 'auto',

'max_depth': 80,

'bootstrap': True

}

**Gradient boost**

{ 'n_estimators': 1000,
'min_samples_split': 6,
'min_samples_leaf': 8,
'max_features': 'log2',
'max_depth': 20,
'learning_rate': 0.01
}

**Neural network**

{ 'n_neurons': 15, 'n_hidden': 3, 'learning_rate': 0.0001}

**SVM**

{ 'C': 7.924145688620425,
'gamma': 0.14645041271999773
}

**LOSO**

***KNN***

*{'weights': 'distance',*

*'n_neighbors': 73,*

*'metric': 'manhattan'}*

***Random Forest***

*{'n_estimators': 1000,*

*'min_samples_split': 10,*

*'min_samples_leaf': 4,*

*'max_features': 'auto',*

*'max_depth': 20,*

*'bootstrap': True}*

***Gradient boost***

*{ 'n_estimators': 1000,*
*'min_samples_split': 6,*

*'min_samples_leaf': 2,*

*'max_features': 'log2',*

*'max_depth': 3,*

*'learning_rate': 0.01*

*}*

***Neural network***

*{ 'n_neurons': 65,*
*'n_hidden': 2,*
*'learning_rate': 0.0001}*

**SVM**

{ 'C': 7.924145688620425,
'gamma': 0.14645041271999773}

## Appendix 3.2 LOSO results

| Model | Activity | Predicted (METs) | True (METs) | MAPE | RMSE | CCC (95% CI) |
|---|---|---|---|---|---|---|
| **AG Gradient Boost** | ADL | 2.82 ± 0.91 | 2.56 ± 0.89 | 22.93156 | 0.723384 | 0.69 (0.66, 0.72) |
| **AG Gradient Boost** | Cycling | 4.7 ± 1.29 | 4.82 ± 1.59 | 16.53547 | 1.065576 | 0.73 (0.7, 0.76) |
| **AG Gradient Boost** | Elliptical | 6.75 ± 1.52 | 7.04 ± 2.13 | 15.06007 | 1.517127 | 0.67 (0.61, 0.72) |
| **AG Gradient Boost** | Rowing | 6.35 ± 1.55 | 6.51 ± 2.04 | 14.33669 | 1.243145 | 0.76 (0.72, 0.8) |
| **AG Gradient Boost** | Running | 8.25 ± 1.3 | 8.52 ± 1.66 | 13.55442 | 1.41443 | 0.56 (0.5, 0.61) |
| **AG Gradient Boost** | Sedentary | 1.37 ± 0.34 | 1.3 ± 0.34 | 19.84489 | 0.374868 | 0.4 (0.35, 0.45) |
| **AG Gradient Boost** | Transitional | 3.06 ± 1.83 | 2.99 ± 1.99 | 19.96356 | 0.772233 | 0.92 (0.91, 0.93) |
| **AG Gradient Boost** | Walking | 4.2 ± 0.75 | 4.22 ± 0.99 | 14.45745 | 0.775034 | 0.61 (0.57, 0.65) |
| **AG Neural Network** | ADL | 2.69 ± 1.07 | 2.56 ± 0.89 | 26.0991 | 0.876418 | 0.61 (0.57, 0.64) |
| **AG Neural Network** | Cycling | 4.75 ± 1.64 | 4.82 ± 1.59 | 20.72147 | 1.29029 | 0.68 (0.64, 0.71) |
| **AG Neural Network** | Elliptical | 6.9 ± 1.81 | 7.04 ± 2.13 | 18.07532 | 1.618295 | 0.66 (0.6, 0.72) |
| **AG Neural Network** | Rowing | 6.26 ± 1.9 | 6.51 ± 2.04 | 16.74395 | 1.391849 | 0.75 (0.7, 0.8) |
| **AG Neural Network** | Running | 8.41 ± 1.82 | 8.52 ± 1.66 | 16.72177 | 1.822038 | 0.45 (0.38, 0.52) |
| **AG Neural Network** | Sedentary | 1.34 ± 0.41 | 1.3 ± 0.34 | 24.361 | 0.461023 | 0.25 (0.19, 0.3) |
| **AG Neural Network** | Transitional | 3.1 ± 2.09 | 2.99 ± 1.99 | 25.22219 | 1.064249 | 0.86 (0.84, 0.88) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **AG Neural Network** | Walking | 4.26 ± 1.12 | 4.22 ± 0.99 | 17.713 48 | 0.9744 02 | 0.57 (0.53, 0.62) |
| **AG Random Forest** | ADL | 2.86 ± 0.92 | 2.56 ± 0.89 | 24.160 11 | 0.7428 28 | 0.68 (0.65, 0.71) |
| **AG Random Forest** | Cycling | 4.71 ± 1.24 | 4.82 ± 1.59 | 16.715 48 | 1.0643 88 | 0.72 (0.69, 0.75) |
| **AG Random Forest** | Elliptical | 6.76 ± 1.47 | 7.04 ± 2.13 | 15.154 74 | 1.5161 9 | 0.66 (0.6, 0.71) |
| **AG Random Forest** | Rowing | 6.3 ± 1.49 | 6.51 ± 2.04 | 14.307 61 | 1.2529 87 | 0.75 (0.71, 0.79) |
| **AG Random Forest** | Running | 8.22 ± 1.24 | 8.52 ± 1.66 | 13.585 71 | 1.4242 02 | 0.54 (0.48, 0.59) |
| **AG Random Forest** | Sedentary | 1.38 ± 0.35 | 1.3 ± 0.34 | 20.308 17 | 0.3915 11 | 0.38 (0.33, 0.43) |
| **AG Random Forest** | Transitional | 3.09 ± 1.79 | 2.99 ± 1.99 | 21.006 82 | 0.7889 97 | 0.91 (0.9, 0.92) |
| **AG Random Forest** | Walking | 4.2 ± 0.68 | 4.22 ± 0.99 | 14.541 04 | 0.7691 56 | 0.59 (0.55, 0.63) |
| **FB Gradient Boost** | ADL | 3.27 ± 1.1 | 2.55 ± 0.89 | 41.943 53 | 1.2231 18 | 0.41 (0.37, 0.45) |
| **FB Gradient Boost** | Cycling | 3.97 ± 1.29 | 4.76 ± 1.58 | 25.296 85 | 1.6741 85 | 0.42 (0.37, 0.46) |
| **FB Gradient Boost** | Elliptical | 6.34 ± 1.51 | 7.01 ± 2.16 | 21.489 92 | 1.9512 74 | 0.48 (0.4, 0.56) |
| **FB Gradient Boost** | Rowing | 4.95 ± 1.24 | 6.49 ± 2.05 | 25.568 61 | 2.3860 1 | 0.3 (0.23, 0.36) |
| **FB Gradient Boost** | Running | 8.3 ± 1.31 | 8.51 ± 1.67 | 14.463 02 | 1.5032 26 | 0.5 (0.44, 0.56) |
| **FB Gradient Boost** | Sedentary | 1.55 ± 0.53 | 1.29 ± 0.34 | 34.738 25 | 0.6533 84 | 0.08 (0.03, 0.13) |
| **FB Gradient Boost** | Transitional | 3.47 ± 1.61 | 2.98 ± 1.99 | 46.036 61 | 1.2861 93 | 0.76 (0.72, 0.79) |
| **FB Gradient** | Walking | 4.52 ± 0.97 | 4.21 ± | 19.011 | 0.9987 | 0.51 (0.46, |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Boost** | | | 0.99 | 7 | 51 | 0.55) |
| **FB Neural Network** | ADL | 3.21 ± 1.27 | 2.55 ± 0.89 | 43.441 56 | 1.3361 11 | 0.37 (0.33, 0.41) |
| **FB Neural Network** | Cycling | 3.89 ± 1.41 | 4.76 ± 1.58 | 27.878 47 | 1.8050 71 | 0.38 (0.33, 0.43) |
| **FB Neural Network** | Elliptical | 6.32 ± 1.81 | 7.01 ± 2.16 | 24.550 78 | 2.0761 1 | 0.49 (0.4, 0.56) |
| **FB Neural Network** | Rowing | 5.23 ± 1.43 | 6.49 ± 2.05 | 24.891 63 | 2.2580 44 | 0.35 (0.27, 0.43) |
| **FB Neural Network** | Running | 8.31 ± 1.52 | 8.51 ± 1.67 | 15.942 36 | 1.6371 89 | 0.48 (0.41, 0.54) |
| **FB Neural Network** | Sedentary | 1.53 ± 0.52 | 1.29 ± 0.34 | 37.521 9 | 0.6580 61 | 0.01 (-0.03, 0.06) |
| **FB Neural Network** | Transitional | 3.42 ± 1.73 | 2.98 ± 1.99 | 48.180 93 | 1.4073 43 | 0.72 (0.68, 0.76) |
| **FB Neural Network** | Walking | 4.55 ± 1.1 | 4.21 ± 0.99 | 21.173 96 | 1.0791 41 | 0.49 (0.44, 0.54) |
| **FB Random Forest** | ADL | 3.31 ± 1.1 | 2.55 ± 0.89 | 42.512 39 | 1.2275 45 | 0.41 (0.37, 0.45) |
| **FB Random Forest** | Cycling | 3.95 ± 1.25 | 4.76 ± 1.58 | 24.504 08 | 1.6615 26 | 0.42 (0.37, 0.46) |
| **FB Random Forest** | Elliptical | 6.29 ± 1.38 | 7.01 ± 2.16 | 20.527 28 | 1.9188 14 | 0.48 (0.4, 0.55) |
| **FB Random Forest** | Rowing | 4.91 ± 1.13 | 6.49 ± 2.05 | 25.313 52 | 2.3971 05 | 0.28 (0.21, 0.34) |
| **FB Random Forest** | Running | 8.21 ± 1.24 | 8.51 ± 1.67 | 14.328 19 | 1.4890 55 | 0.5 (0.43, 0.55) |
| **FB Random Forest** | Sedentary | 1.54 ± 0.49 | 1.29 ± 0.34 | 33.731 8 | 0.6229 48 | 0.08 (0.03, 0.13) |
| **FB Random Forest** | Transitional | 3.49 ± 1.56 | 2.98 ± 1.99 | 47.271 92 | 1.2937 61 | 0.75 (0.71, 0.78) |
| **FB Random Forest** | Walking | 4.54 ± 0.95 | 4.21 ± 0.99 | 19.177 41 | 0.9928 9 | 0.51 (0.46, 0.55) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **SWA Gradient Boost** | ADL | 2.85 ± 1.04 | 2.56 ± 0.89 | 23.266 55 | 0.8164 61 | 0.66 (0.62, 0.69) |
| **SWA Gradient Boost** | Cycling | 4.66 ± 1.34 | 4.84 ± 1.58 | 16.600 19 | 1.0655 48 | 0.74 (0.71, 0.77) |
| **SWA Gradient Boost** | Elliptical | 6.97 ± 1.7 | 7.13 ± 2.12 | 14.162 44 | 1.3854 26 | 0.74 (0.69, 0.78) |
| **SWA Gradient Boost** | Rowing | 6.08 ± 1.67 | 6.58 ± 2.04 | 14.423 18 | 1.2911 94 | 0.77 (0.72, 0.81) |
| **SWA Gradient Boost** | Running | 8.31 ± 1.35 | 8.54 ± 1.66 | 12.000 7 | 1.2507 24 | 0.66 (0.61, 0.7) |
| **SWA Gradient Boost** | Sedentary | 1.36 ± 0.35 | 1.3 ± 0.34 | 20.339 26 | 0.4150 76 | 0.29 (0.23, 0.34) |
| **SWA Gradient Boost** | Transitional | 3.14 ± 1.88 | 3 ± 1.99 | 21.871 51 | 0.8229 31 | 0.91 (0.89, 0.92) |
| **SWA Gradient Boost** | Walking | 4.26 ± 0.88 | 4.24 ± 0.99 | 12.891 42 | 0.6935 5 | 0.73 (0.69, 0.76) |
| **SWA Manufacturer** | ADL | 3.73 ± 2.05 | 2.57 ± 0.89 | 66.517 31 | 2.2817 7 | 0.18 (0.14, 0.21) |
| **SWA Manufacturer** | Cycling | 3.31 ± 1.88 | 4.83 ± 1.58 | 38.762 46 | 2.1830 26 | 0.43 (0.39, 0.47) |
| **SWA Manufacturer** | Elliptical | 6 ± 1.57 | 7.13 ± 2.12 | 24.407 59 | 2.6749 19 | 0.13 (0.04, 0.22) |
| **SWA Manufacturer** | Rowing | 6.14 ± 2.1 | 6.58 ± 2.04 | 33.958 17 | 2.6839 81 | 0.18 (0.06, 0.28) |
| **SWA Manufacturer** | Running | 8.11 ± 1.69 | 8.54 ± 1.66 | 21.949 49 | 2.2775 58 | 0.1 (0.02, 0.18) |
| **SWA Manufacturer** | Sedentary | 1.22 ± 0.27 | 1.3 ± 0.34 | 21.438 14 | 0.4043 44 | 0.15 (0.09, 0.21) |
| **SWA Manufacturer** | Transitional | 2.77 ± 1.79 | 3 ± 1.99 | 31.170 51 | 1.4589 13 | 0.7 (0.66, 0.74) |
| **SWA Manufacturer** | Walking | 3.92 ± 0.85 | 4.24 ± 0.99 | 21.636 17 | 1.1928 05 | 0.21 (0.15, 0.27) |
| **SWA Neural** | ADL | 2.78 ± 1.07 | 2.56 ± | 24.256 | 0.8185 | 0.66 (0.63, |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Network** | | | 0.89 | 3 | 16 | 0.69) |
| **SWA Neural Network** | Cycling | 4.65 ± 1.44 | 4.84 ± 1.58 | 17.295 35 | 1.1366 08 | 0.72 (0.69, 0.75) |
| **SWA Neural Network** | Elliptical | 7.05 ± 1.8 | 7.13 ± 2.12 | 15.446 56 | 1.4636 28 | 0.72 (0.66, 0.77) |
| **SWA Neural Network** | Rowing | 6.25 ± 1.83 | 6.58 ± 2.04 | 15.773 76 | 1.3030 15 | 0.78 (0.73, 0.82) |
| **SWA Neural Network** | Running | 8.39 ± 1.51 | 8.54 ± 1.66 | 11.769 15 | 1.2617 03 | 0.68 (0.64, 0.73) |
| **SWA Neural Network** | Sedentary | 1.36 ± 0.44 | 1.3 ± 0.34 | 24.066 72 | 0.5091 83 | 0.16 (0.1, 0.22) |
| **SWA Neural Network** | Transitional | 3.13 ± 2.02 | 3 ± 1.99 | 25.215 73 | 0.9146 06 | 0.9 (0.88, 0.91) |
| **SWA Neural Network** | Walking | 4.26 ± 0.99 | 4.24 ± 0.99 | 14.552 55 | 0.7612 54 | 0.71 (0.67, 0.74) |
| **SWA Random Forest** | ADL | 2.93 ± 1.06 | 2.56 ± 0.89 | 24.877 31 | 0.8646 05 | 0.63 (0.6, 0.67) |
| **SWA Random Forest** | Cycling | 4.66 ± 1.29 | 4.84 ± 1.58 | 16.062 89 | 1.0465 59 | 0.74 (0.71, 0.77) |
| **SWA Random Forest** | Elliptical | 6.91 ± 1.61 | 7.13 ± 2.12 | 14.156 81 | 1.3842 97 | 0.73 (0.68, 0.77) |
| **SWA Random Forest** | Rowing | 6.02 ± 1.64 | 6.58 ± 2.04 | 14.404 34 | 1.3092 21 | 0.76 (0.71, 0.8) |
| **SWA Random Forest** | Running | 8.28 ± 1.26 | 8.54 ± 1.66 | 12.275 91 | 1.2563 8 | 0.64 (0.59, 0.68) |
| **SWA Random Forest** | Sedentary | 1.37 ± 0.36 | 1.3 ± 0.34 | 20.661 21 | 0.4259 22 | 0.26 (0.21, 0.32) |
| **SWA Random Forest** | Transitional | 3.16 ± 1.83 | 3 ± 1.99 | 22.978 23 | 0.8220 99 | 0.91 (0.89, 0.92) |
| **SWA Random Forest** | Walking | 4.26 ± 0.83 | 4.24 ± 0.99 | 12.874 53 | 0.6791 99 | 0.72 (0.69, 0.75) |

LOSO results for each of the regression models by the activity types. Abbreviations: Fitbit (FB), AG = ActiGraph (AG), SenseWear (SWA). Root mean squared error (RMSE), Mean absolute percentage error (MAPE), concordance correlation coefficient (CCC).

# Appendix 4.1 Distributions used in the hierarchical modelling approaches

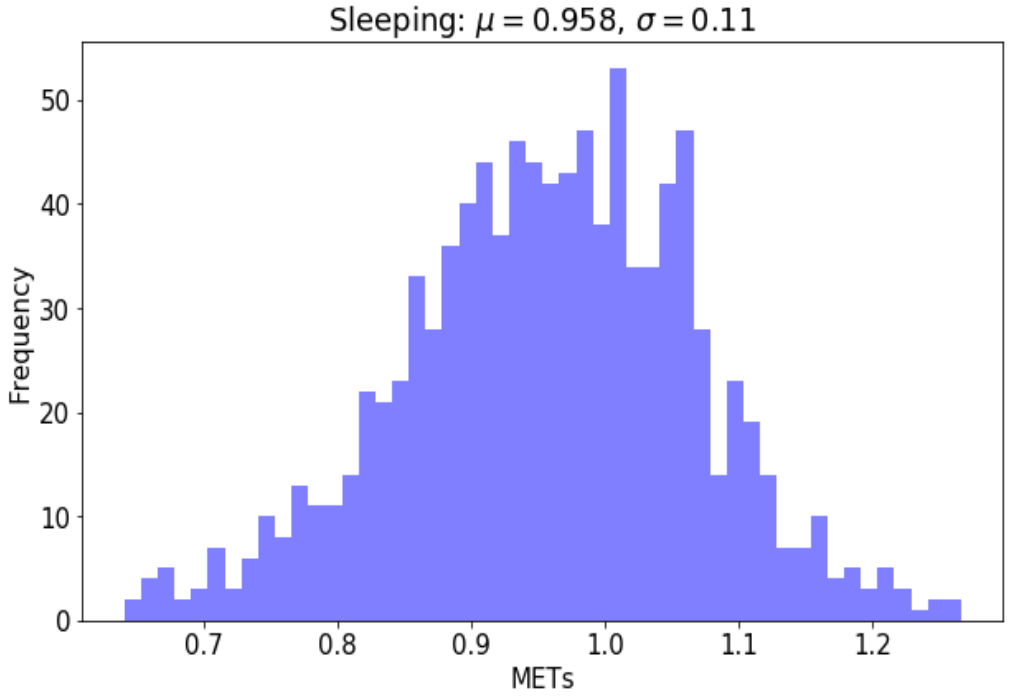



**Figure 1.** A histogram of 1000 draws from the METs distribution used for minutes where the subject was sleeping.

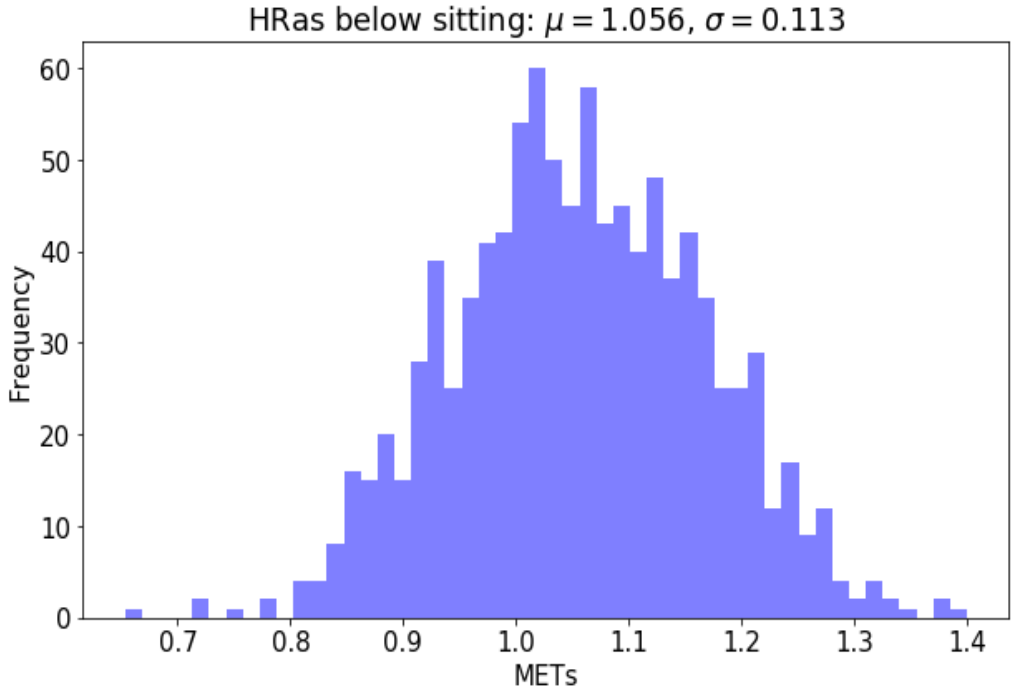



**Figure 2.** A histogram of 1000 draws from the METs distribution used for minutes in which measured heart rate was lower than the sitting heart rate.

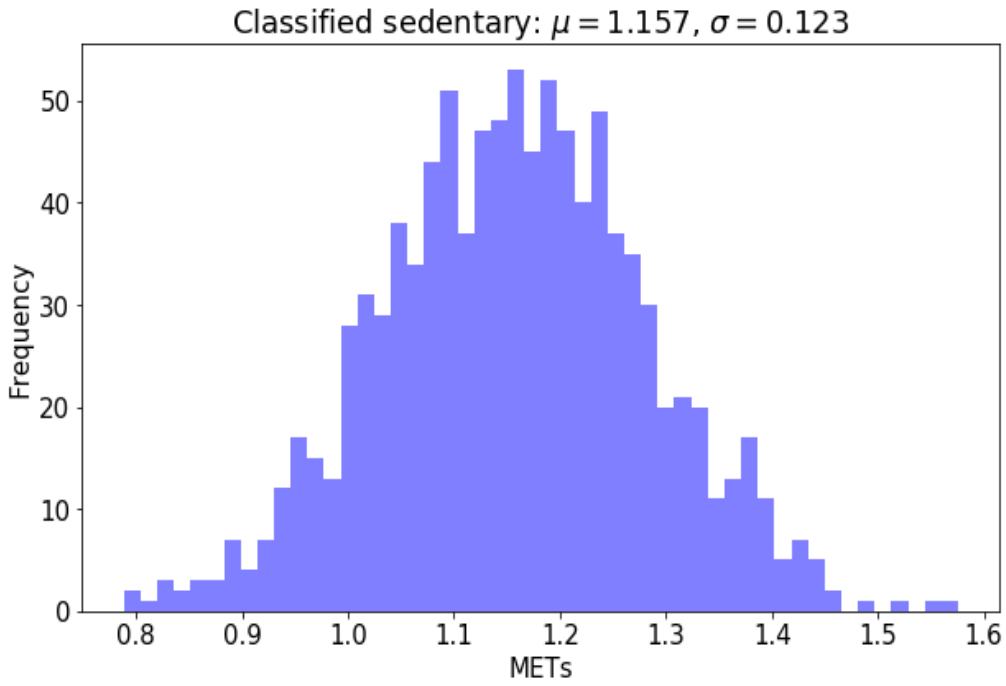



**Figure 3.** A histogram of 1000 draws from the METs distribution used for minutes classified as sedentary by the classification algorithm.

# Appendix 4.2 Energy intake estimates

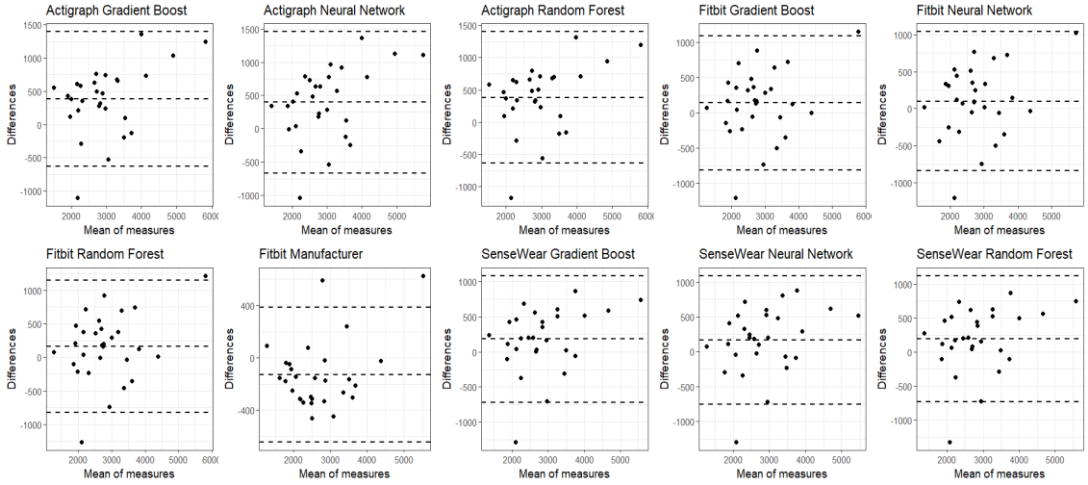



**Figure 1.** Bland-Altman plots detailing the differences between the respective models and the SenseWear armband for energy intake.

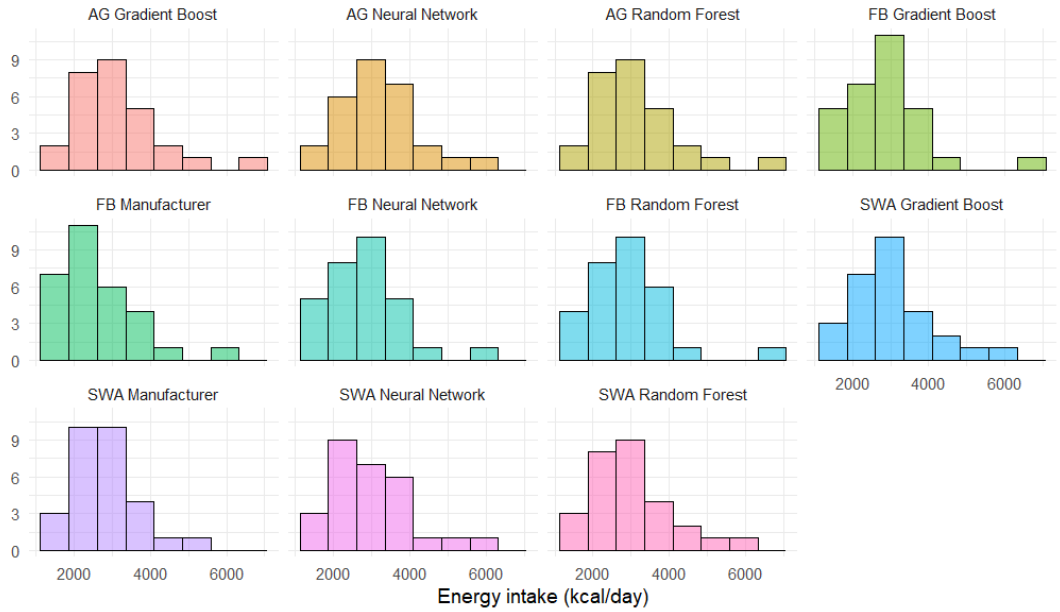



**Figure 2** Histograms detailing the distribution of energy intake for each of the predictive models, the SenseWear and Fitbit manufacturer estimates.

# Appendix 5.1 Visualisations of $\Delta EI$ estimates

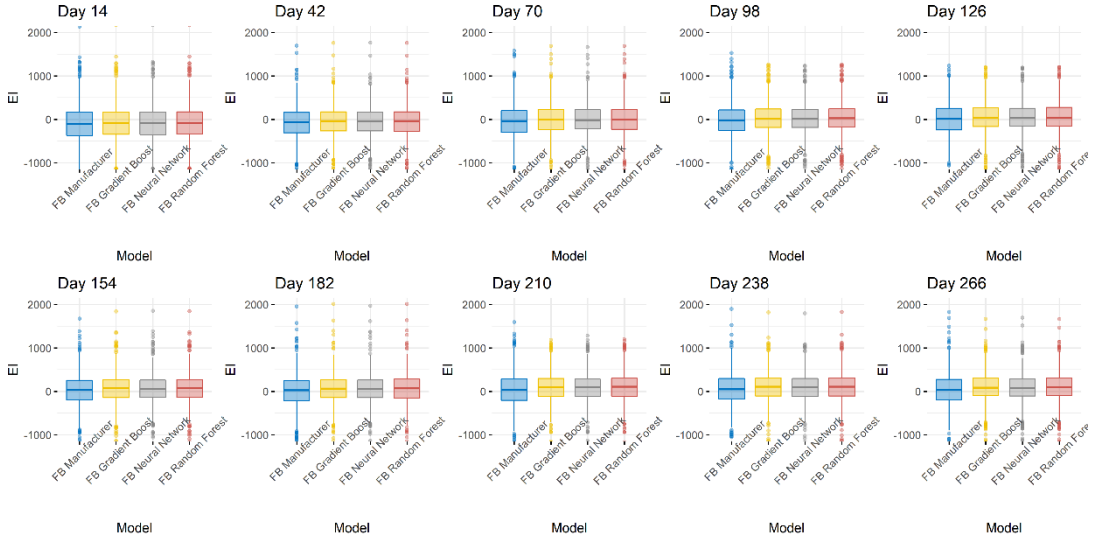



**Figure 1.** Boxplots demonstrating the variability in ΔEI for each of the models for the second half of the study

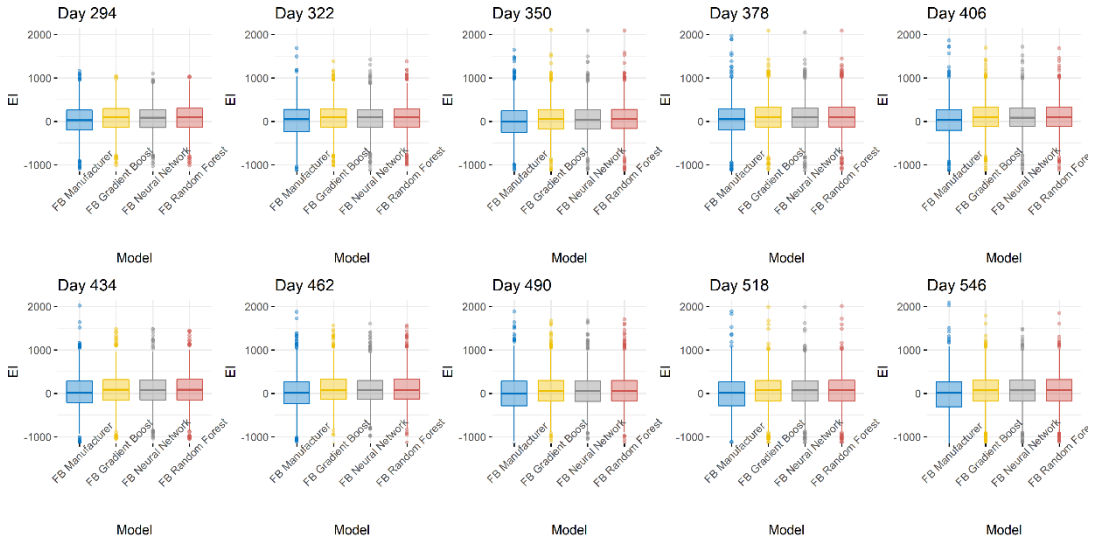



**Figure 2.** Boxplots demonstrating the variability in ΔEI for each of the models for the second half of the study

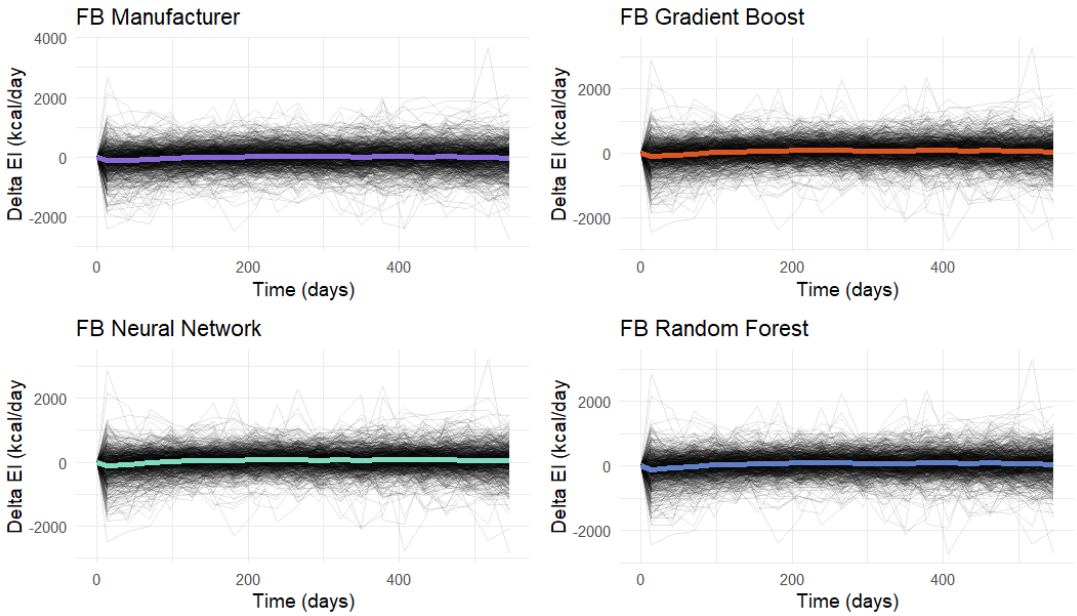



**Figure 3.** Line plots demonstrating ΔEI for each of the models. The solid line represents the average ΔEI estimate for the model and the black lines represent individual subjects.