UNIVERSITY OF LEEDS

Temporal Graph-based Convolutional Neural Networks for Electronic Health Records

Zoe Louise Hancox

Submitted in accordance with the requirements for the degree of MSc. and PhD. Artificial Intelligence for Medical Diagnosis and Care

The University of Leeds

Faculty of Engineering

School of Computing

February 2025

Intellectual Property and Publication Statements

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Signed 

Acknowledgements

First and foremost, I would like to express my sincere gratitude to the Centre for Doctoral Training (CDT) for AI in Medical Diagnosis and Care and its directors for giving me the opportunity to undertake this research. Their support has been invaluable in shaping my academic journey and allowing me to contribute to the field of AI-driven healthcare. I am also grateful for the financial support I received through EPSRC funding (Grant No. EP/S024336/1), which made this work possible.

I would like to extend my appreciation to my clinical supervisors (Sarah Kingsbury, Philip Conaghan, and Andy Clegg) for their expertise, guidance, and patience. Their insights not only helped me better understand complex clinical problems but also taught me how to frame research in a way that is meaningful and applicable to real-world healthcare challenges.

A huge thank you to my primary supervisor Sam Relton, whose unwavering support and kindness have been truly invaluable. Without his guidance, patience, and encouragement, I am not sure I would have made it to the end of this PhD. He is one of the most inspiring and dedicated mentors I have ever had the privilege to work with, always willing to help, endlessly generous with his time, and never making me feel like an inconvenience. His mentorship has not only shaped this research but has also had a profound impact on my academic and personal growth.

A special thank you to Allan Pang, who generously volunteered his time to serve as a second reviewer for the systematic literature review included in this thesis. His patience and diligence in screening hundreds of papers, played a crucial role in ensuring the robustness and quality of this work. I am truly grateful for his support, attention to detail, and perseverance throughout the process.

A heartfelt thank you to my office buddies (and minions) for their unwavering support, camaraderie, and the countless pieces of invaluable advice they've thrown my way. The collaborative

and uplifting environment they've fostered has made this journey all the more enjoyable. Their kindness, encouragement, and sheer brilliance have been truly inspiring, so much so that they motivated me to apply for a Research Fellowship within the same department. I am excited for the opportunity to continue working alongside such amazing and talented colleagues on groundbreaking research.

I am incredibly grateful to my family for their love and encouragement throughout this PhD journey. Their support has been the foundation that kept me going through the challenges, late nights, and moments of doubt. Thank you for always believing in me, even when I struggled to believe in myself. Thank you for always being there to lift my spirits, make me laugh, and remind me that there is life outside of research. To my parents, thank you for your endless support and for instilling in me the values of perseverance, curiosity, and hard work. Your encouragement has been my strength, and I could not have done this without you.

And to my partner, Alex, I cannot thank you enough for your patience, understanding, and unwavering belief in me. Your support has meant the world to me, and I am endlessly grateful for the way you've been by my side through every high and low of this journey. Thank you for dragging me away from the computer when I needed rest, for always looking after me, and for being my steady source of comfort and reassurance. Your willingness to listen to my worries and rants (no matter how repetitive) has been a lifeline, and your reminders that things aren't as daunting as they seem have helped me regain perspective time and time again. I couldn't have done this without you.

Finally, this research would not have been possible without the data provided by patients and collected by the NHS as part of their care and support. I deeply appreciate every individual who has chosen to remain opted in for data sharing, enabling researchers like myself to work with anonymised data to develop models that aim to improve health outcome risk prediction. Their contribution to research is invaluable, and I hope that this work will, in some small way, contribute to advancing patient care and outcomes.

Project Approval

Approval for the study was obtained from the School of Medicine Research Ethics Committee (SoMREC) at the University of Leeds (reference: SoMREC/13/079), and the Research Project Committee at ResearchOne (project number: 201428378A).

Abstract

Graph theory offers a powerful framework for using the relational dependencies in Electronic Health Records (EHRs) to enhance machine learning (ML) predictions of health outcomes and diagnoses. This thesis explores and advances graph-based ML approaches, with applications for the prediction of future hip and knee replacement risk.

A systematic literature review identified 832 studies, with 18 using patient-level graph representations of EHRs for predicting health outcomes. This review showed that current graph-based EHR models have limited clinical applicability due to high risk of bias.

A novel Temporal Graph-Based Convolutional Neural Network (TG-CNN) model was developed. Initially applied to student dropout prediction in online courses, this approach demonstrated state-of-the-art performance. Extending this method to medical data, TG-CNNs were applied to predict hip and knee replacement risks, one and five years in advance. Temporal graphs, constructed from primary care event codes from EHRs, captured temporal relationships between symptoms, diagnoses, and prescriptions. Models achieved AUROC values up to 0.967 for hip replacement and 0.955 for knee replacement.

To improve model interpretability, four explainable methods were explored, including gradient-based and feature-mapping approaches. These methods provided visual insights into TG-CNN predictions, highlighting the influence of key EHR features such as prescriptions. While clinicians found these visualisations informative, further simplification is needed to support real-world clinical decision-making.

This thesis demonstrates that graph-based representations improve the predictive performance and interpretability of ML models in healthcare. The TG-CNN model offers the potential to enhance patient care and management through earlier and more accurate predictions. Future work should focus on improving model explainability and translation into clinical practice.

Publications Arising from Thesis

This research has been carried out by myself (Zoe Hancox (ZH)) and my supervisory team which includes Samuel Relton (SR), Philip Conaghan (PC), Sarah Kingsbury (SK), and Andrew Clegg (AC). My own contributions, fully and explicitly indicated in the thesis, have been to design and conduct the methodology for each publication, draft the initial versions of the papers, and edit the final publications. The other authors provided feedback and/or served as supervisors for all the publications, except for the systematic review. In that case, the second author Allan Pang (AP) also contributed to the results, specifically by screening papers, extracting data, and conducting a risk of bias analysis independently to my risk of bias analysis to ensure robustness. The other members of the group and their contributions are outlined as follows for each publication chapter:

Chapter 3 (Systematic literature review): 'A Systematic Review of Networks for Prognostic Prediction of Health Outcomes and Diagnostic Prediction of Health Conditions within Electronic Health Records', authors: ZH, AP, PC, SK, AC, SR (published in the Artificial Intelligence in Medicine journal). In the publication arising from this chapter, ZH was responsible for formal analysis and investigation. SR served as the mediator. The methodology was developed collaboratively by ZH, SR, SK, PC, and AC. ZH and AP carried out the paper screening and data extraction. ZH acted as the review guarantor, ensuring the integrity of the review process. Supervision was provided by SR, PC, SK, and AC. The original draft was written by ZH and AP, while all authors, including SR, SK, PC, and AC, contributed to the review and editing of the manuscript. This thesis chapter includes extra information on top of the information provided in the corresponding publication, this includes electronic health record sample size statistics from the studies and additional conclusions and future directions relative to the thesis, alongside Section 3.5.3.

Chapter 4 (Model methodology and development): 'Temporal Graph-based CNNs (TG-CNNs) for Online Course Dropout Prediction', authors: ZH, SR (ISMIS 2022 conference). Chapter 4 includes additional information in Sections 4.2.2, 4.2.4, 4.2.5 and 4.2.6 and this chapter builds on the conclusions relevant to the research questions.

Chapter 5 (Model methodology for clinical application): 'Developing the Temporal Graph Convolutional Neural Network Model to Predict Hip Replacement using Electronic Health Records', authors: ZH, SK, AC, PC, SR (ICMLA 2024 conference). Chapter 5 includes additional information on model components in the discussion and more details in the limitations of this thesis chapter compared to the corresponding publication.

Chapter 6 (Comparing hip and knee prediction models): 'Primary Care Prediction of Hip and Knee Replacement 1-5 Years in Advance Using Temporal Graph-based Convolutional Neural Networks (TG-CNNs)', authors: ZH, SK, PC, AC, SR (under review in the Rheumatology Oxford Journal). In addition to the submitted paper, Chapter 6 of this thesis includes dedicated sections on limitations, future work, and conclusions (Sections 6.4.2, 6.4.3, and 6.5). Unlike the submitted paper, the literature review and results are integrated directly into the chapter rather than the appendices. The related work section (Section 6.4.1) expands on secondary care predictor models in addition to primary care predictor models, providing a more comprehensive discussion.

Chapter 7 (Explainable graph-based models): 'Explainable Temporal Graph-Based CNNs for Predicting Hip Replacement Risk using EHR Data', authors: ZH, David Wong, SK, PC, AC, SR (presented as poster at ML4H 2024 conference and submitted to the Open Journal of the American Medical Informatics Association (JAMIA Open)). David Wong contributed to reviewing and editing the article for the journal. This thesis chapter contains additional figures describing each of the four explainability methodologies in more detail compared to the corresponding research paper submission. Subgraph frequency analysis also included three additional research questions, and additional figures and content in Section 7.4.5.

Contents

List of Figures

List of Tables

Abbreviations

ACT MOOC Accessible Culture & Training Massive Open Online Course.

AI Artificial Intelligence.

AMI Acute Myocardial Infarction.

AN-SNAP Australian National Subacute and Non-Acute Patient Classification.

ANN Artificial Neural Network.

AoM Acute Otitis Media.

ATC Anatomical Therapeutic Chemical Classification.

AUC Area Under the Curve.

AUPRC Area Under Precision-Recall Curve.

AUROC Area Under the Receiver Operating Characteristic.

BERT Bi-directional Encoder Representation from Transformers.

BI-LSTM Bi-Directional Long Short-Term Memory.

BMI Body Mass Index.

BNF British National Formulary.

CAM Class Activation Mapping.

CAP Community Acquired Pneumonia.

CHARMS CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies.

CKD Chronic Kidney Disease.

CNN Convolutional Neural Network.

COPD Chronic Obstructive Pulmonary Disease.

CPRD Clinical Practice Research Datalink.

CPT Current Procedural Terminology.

CTV3 Clinical Terms Version 3.

CVD Cardiovascular Disease.

DAG Directed Acyclic Graph.

DM Diabetes Mellitus.

DT Decision Tree.

EDB Edge Detection Bias.

eFI Electronic Frailty Index.

EHR Electronic Health Record.

eICU eICU Collaborative Research Database.

FCL Fully Connected Layer.

FN False Negative.

FP False Positive.

FPR False Positive Rate.

GCN Graph Convolutional Network.

GCT Graph Convolutional Transformer.

GDPR General Data Protection Regulation.

GNN Graph Neural Network.

GP General Practitioner.

Grad-CAM Gradient-based Class Activation Mapping.

GRU Gated Recurrent Unit.

HF Heart Failure.

HTN Hypertension.

ICD International Classification of Diseases.

ICU Intensive Care Unit.

IMD Index of Multiple Deprivation.

IVF In Vitro Fertilisation.

KL Kellgren-Lawrence.

LIME Local Interpretable Model-agnostic Explanations.

LOINC Logical Observation Identifiers Names and Codes.

LR Logistic Regression.

LSTM Long Short-Term Memory.

MIMIC Medical Information Mart for Intensive Care.

ML Machine Learning.

MOOC Massive Open Online Course.

MRI Magnetic Resonance Imaging.

MSK Musculoskeletal.

NHIRD National Health Insurance Research Database.

NHS National Health Service.

NLP Natural Language Processing.

NOA Non-opioid Analgesic.

NPV Negative Predictive Value.

NSAID Non-Steroid Anti-Inflammatory Drug.

OA Osteoarthritis.

PPIE Patient and Public Involvement and Engagement.

PPV Positive Predictive Value.

PRISMA Preferred Reporting Items for Systematic Review and Meta-Analysis.

PROBAST Prediction model Risk Of Bias Assessment Tool.

RA Rheumatoid Arthritis.

ReLU Rectified Linear Unit.

RETAIN REverse Time AttentIoN.

RF Random Forest.

RNN Recurrent Neural Network.

RoB Risk of Bias.

SHAP SHapley Additive exPlanations.

SNOMED-CT Systematized Nomenclature of Medicine – Clinical Terms.

SVM Support Vector Machine.

TG-CNN Temporal Graph-Based Convolutional Neural Network.

TGN Temporal Graph Network.

TN True Negative.

TP True Positive.

TPP The Phoenix Partnership.

TPR True Positive Rate.

TRE Trusted Research Environment.

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

UK United Kingdom.

UMLS Unified Medical Language System.

URTI Upper Respiratory Tract Infection.

UTI Urinary Tract Infection.

WOMAC Western Ontario and McMaster Universities Osteoarthritis Index.

Chapter 1

Introduction

1.1 Background

The evolving landscape of healthcare faces unprecedented challenges as advances in medicine, increasingly complex patient treatment pathways, and ageing populations place growing strains on health systems. The integration of predictive algorithms into healthcare offers techniques to reduce clinicians' cognitive burdens, streamline decision-making, and enhance patient outcomes by reducing waiting times for care [1]. The government are encouraging the growth of Artificial Intelligence (AI) within the United Kingdom (UK) to help speed up assessment and diagnosis in healthcare, particularly with patient scans which are becoming increasingly in demand leading to longer waiting times [2].

Many clinical prediction models rely on summary data derived from EHR codes. These codes contain both temporal information (e.g., the timing and sequence of medical events) and structural information (e.g., patient demographics, diagnoses, and procedure codes). However, due to model architecture and preprocessing, temporal features are often discarded or lost during training [3]. Many models simplify or aggregate data to reduce complexity, stripping away valuable contextual information that could be critical for accurate predictions.

Graph-based representations of EHRs offer a way to preserve these complexities, capturing temporal and relational dependencies to improve Machine Learning (ML) predictions. Such advancements hold particular promise for chronic, progressive conditions where early intervention can significantly alter disease trajectories. One such condition is Osteoarthritis (OA), a

leading cause of disability worldwide. With OA cases rising and joint replacement surgeries placing increasing strain on healthcare systems, predictive insights could play a crucial role in enabling earlier interventions, such as physiotherapy or medication, to slow disease progression. By improving care efficiency, optimising resource allocation, and enhancing patient quality of life, these predictive models address clinical and systemic challenges in managing OA [4, 5].

1.2 Motivation

Conditions like hip OA, which affects physical activity and quality of life, are increasingly common and resource-intensive to manage, which regularly result in hip replacement surgery being required [6]. In 2023, 108,558 primary total hip replacements were were recorded by the National Joint Registry in the UK, of which 92% were due to OA [7].

The burden of Musculoskeletal (MSK) conditions, particularly OA, highlights the critical need for predictive tools. Predicting the need for such procedures in advance could significantly enhance healthcare planning and delivery. For example, identifying at-risk patients early would enable timely interventions, such as physiotherapy or lifestyle adjustments, reducing the need for surgery and alleviating patient suffering [8]. This is particularly crucial as ageing demographics and increasing obesity rates exacerbate the prevalence of joint degeneration and the associated strain on healthcare systems.

Furthermore, the adoption of explainable ML models addresses a critical barrier to clinical trust [9]. Predictive algorithms often function as "black-box" systems, offering little transparency about their decision-making processes. In healthcare, this lack of explainability not only undermines trust but also raises ethical concerns, particularly regarding patient safety and fair decision-making. Transparent and interpretable models empower clinicians to understand and validate predictions [10], enabling better communication with patients and compliance with regulations such as the General Data Protection Regulation (GDPR) [11]. As such, explainability is not just a technical requirement but a vital feature to ensure equitable, safe, and patient-centred care.

By integrating predictive algorithms into healthcare systems, resource allocation can be optimised, enabling more strategic planning at local and national levels. This includes reducing unnecessary procedures, mitigating costs, and improving patient outcomes, aligning with public

health goals to enhance the efficiency and equity of care delivery.

1.3 Problem Statement

Predicting long-term health outcomes is essential for effective healthcare planning, early intervention, and personalised treatment strategies. However, accurately forecasting health outcomes in advance remains a significant challenge due to the complex, dynamic, and interconnected nature of health data. Traditional predictive models in healthcare usually work with either static data (such as a patient's age, gender, or diagnosis at a single point in time) or basic time-series data (like repeated measurements of blood pressure or heart rate over time). However, these approaches often fall short because they do not account for the complex and dynamic relationships between different health indicators [12]. For example, how one indicator, such as blood sugar level, might influence another, like blood pressure, over time, or how these relationships change in parallel as a patient's condition evolves. This inability to capture such interdependencies and evolving patterns can limit the model's ability to make accurate and contextually rich predictions, especially in patients with complex EHRs.

Existing methods also struggle with integrating the relational structure of health data, such as the way different health metrics and demographic factors interact over time, which is critical for understanding and predicting health outcomes.

This research addresses the critical need for a predictive model that can use temporal and relational data to more accurately forecast hip and knee replacement risk over extended periods. To do this the Temporal Graph-Based Convolutional Neural Network (TG-CNN) model is introduced which uses EHR data, including the elapsed time between General Practitioner (GP) visits, to predict individual patient outcomes in advance. By developing a TG-CNN model, this thesis aims to fill the gap in predictive healthcare analytics, enabling a more comprehensive and interpretable approach to long-term health outcome prediction.

1.4 Objectives and Research Questions

The objective of this thesis is to determine how TG-CNN models can be used for healthcare predictions using EHR data represented as graphs, where Clinical Codes are the nodes and these are linked by edges using time in between patient visits. The overall research question

can be encapsulated as:

Can incorporating elapsed time between EHR events/clinical codes be used, with the TG-CNN model, to improve predictive power and enable clinicians to make informed decisions more effectively, compared to the current state-of-the-art AI approaches?

With the following sub-questions that aim to be achieved in subsequent chapters:

1. In what ways can graph-based representations and AI improve clinical insights and diagnostic capabilities in MSK research?
2. How are graphs being used on EHRs to predict diagnosis and health outcomes?
   (a) What graph approaches are researchers taking to predict these health outcomes?
   (b) How do these graph approaches compare to each other?
   (c) How are nodes and edges being utilised to perform these tasks?
   (d) How do these approaches compare to other machine learning, artificial intelligence and statistical models?
3. What is an effective architecture for TG-CNN models? Can the TG-CNN be used effectively on simple temporal data to predict binary outcomes?
4. Can the TG-CNN model be used to predict hip and knee replacement risk at one and five-year intervals, and how does its performance compare to existing models in the literature?
5. What methods are useful to apply to TG-CNNs to provide explainability for model predictions to clinicians in a visually understandable way?

1.5 Overview of Datasets

Two datasets are used in this thesis. The first dataset used in this thesis was the Accessible Culture & Training Massive Open Online Course (ACT MOOC) dataset[1] , this is composed of a temporal sequence of timestamped student clickstream actions and subsequent dropout information. The Massive Open Online Course (MOOC) dataset consists of 7,047 users. There are 97 potential clickstream actions a student can take. Further details of this dataset can be found in Chapter 4.

[1]Stanford Network Analysis Project - `https://snap.stanford.edu/data/act-mooc.html`

The second dataset is a health dataset constructed from primary care National Health Service (NHS) EHRs. Specifically, the data is from ResearchOne, which is managed by The Phoenix Partnership (TPP). It comprises of clinical and administrative data from 151,565 patients (aged 40-75) who attended practices in England using SystmOne and who did not use the national opt out option for data sharing. This dataset contains around 2,000 different Clinical Terms Version 3 (CTV3) codes covering clinical details. Patients were eligible for this dataset if their first record of joint pain was clinically coded between April 1st, 1999, and March 31st, 2014 [13]. The ResearchOne data is not distributable under license and is not publicly available. Further details of this dataset are covered in Chapters 5-6.

1.6 Contributions of the Thesis

This thesis aims to provide readers with the knowledge to understand the methodology and motivation behind the TG-CNN model. The TG-CNN model developed during this project is openly available on GitHub[2].

Algorithms were developed to enable the sparse implementation of the 3D Convolutional Neural Network (CNN) layer used within the TG-CNN model for healthcare applications.

This thesis introduces the application of graph-based primary care EHRs for hip and knee replacement prediction. No other studies have used graphs for hip or knee replacement prediction, which enable the use of high dimensional data to improve predictive performance compared to current methods such as Logistic regression and Random Forest Survival models [14, 15].

Not only does this model provide state-of-the-art performance, it also provides an intuitive approach that enables model interpretability and predictive reasoning via interactive graphs. This work includes a small portion of user evaluation, which can be used for future implementation into something with clinical utility.

1.7 Methodology Overview

Research methods involve the development of a codebase constructing the TG-CNN model for experimental analysis. First, the datasets are explored and pre-processed to format them appropriately for feeding into the model. Then the models are trained and tuned based on

[2] `Sparse TGCNN Repository on GitHub: https://github.com/ZoeHancox/Sparse_TGCNN`

various hyperparameters and model layers to optimise the architecture and improve prediction performance. Specifically, this is first performed on the ACT MOOC dataset to predict student dropout and compared against current state-of-the-art models in this area. Then the model is further developed and is trained on EHR data for prediction of hip replacement risk one year in advance, again with further hyperparameter optimisations, ablation studies, existing model comparisons, and performance evaluations. Following that, four models are trained to separately predict hip and knee replacement risk one and five years in advance, observing any changes to model performance. And lastly the model is adjusted to enable explainability, constructing various forms of interactive individual patient graphs using different explainability methods to provide users understanding of model decision reasoning. The interactive graphs are then evaluated asking clinicians to complete a short questionnaire on their experiences with the interactive graphs.

This methodology allows the construction and internal validation of the TG-CNN model to assess the performance in the healthcare domain, predicting hip and knee replacement risk and comparing the model to models in the literature. Whilst considering how the model could be further utilised for clinical decision making and interpretability for transparent modelling.

1.8 Thesis Outline

The first chapter (Chapter 2) of this thesis forms the narrative review in which a brief background about MSK conditions, prediction models, and models used for healthcare prediction are covered. This chapter outlines the current research undertaken to predict MSK related outcomes, whilst also uncovering current use of EHRs and how they could be used further for health outcome predictions.

Chapter 3 covers a systematic review of networks for prognostic prediction of health outcomes and diagnostic prediction of health conditions within EHRs. It looks at the existing literature using graph-based techniques for individual health outcome prediction. It uncovers the key gaps in the literature in this area alongside the current biases present, which help inform the subsequent chapters of this thesis.

Chapter 4 introduces the TG-CNN development and methodology using a simple MOOC clickstream dataset to establish the foundations of the model before application to medical data. It

introduces the sparse 3D CNN layer technique with examples. The chapter then compares 3D CNN sparse and dense implementation of the models.

The TG-CNN model is applied to primary care health data in Chapter 5. Specifically the model is used to predict hip replacement risk one year in advance, in patients who had experienced joint pain.

In Chapter 6 the work is extended to produce models for hip and knee replacement risk one and five years in advance. This chapter looks at comparing the performance of the four models.

Chapter 7, explores various methodologies to make the TG-CNN model explainable. Interactive graphs are developed showing which parts of a patients EHR history contribute to the models prediction decision. Clinical feedback on these interactive graphs is given.

1.9 Expected Impact and Applications

The TG-CNN method is highly versatile and could be used for a variety of scenarios. The method could not only be expanded to other health domains or at different health tiers (secondary or tertiary care), this work has shown that the TG-CNN model is effective for predicting student dropout which could translate to many areas of research.

Project outputs are being used in the DynAIRx[3] project which looks at de-prescribing in patients with polypharmacy, to reduce adverse reactions caused by taking multiple medicines at the same time. The TG-CNN method is being explored to use for better patient prescription of medications calculating risks of hospital admissions and other adverse outcomes for three multimorbidity groups.

Alongside this thesis, an open access GitHub repository was created which not only allows others to train the model to their needs, it also provides the code to perform sparse 3D CNNs, with one degree of freedom, to improve computation speed. Reducing computation costs are vital as the use of AI continues to grow, as the impact of AI changes lives it is important to consider the implications of training and using models on our carbon footprints.

[3] https://www.liverpool.ac.uk/dynairx/

1.10 Summary

This thesis explores the application of TG-CNNs to predict health outcomes, with a focus on forecasting one-year and five-year hip and knee replacement risk. Health outcome prediction is critical for enabling proactive healthcare interventions, especially in managing chronic conditions and anticipating patient needs. Traditional machine learning approaches often overlook the complex temporal dependencies and relational structures inherent in health data, limiting their predictive power over extended periods.

To address this, this research combines the strengths of temporal graph structures and CNNs, to model the evolving relationships and dependencies within longitudinal health data. Temporal graphs allow for a flexible representation of patient EHR data alongside the AI model allowing other health indicators, such as demographic factors to be included besides GP visit data. By embedding this temporal graph information into CNNs, the model can more accurately recognise patterns and correlations across different time frames, improving its predictive accuracy.

The project aims to:

1. Develop and implement the TG-CNN architecture capable of handling high-dimensional, temporally structured health data.
2. Evaluate the model's predictive accuracy in forecasting one-year and five-year hip and knee replacement risk using real-world patient datasets.
3. Assess the interpretability of the model's predictions, providing insights into which factors and pathways most commonly appear in the predictive models.

Through this work, the thesis contributes to advancing predictive analytics in healthcare, providing a novel approach that uses both temporal and relational information within patient data to support early intervention and personalised care planning.

Chapter 2

Narrative Literature Review

2.1 Introduction

Hip and knee replacements are among the most common and successful orthopaedic procedures performed worldwide, alleviating pain and improving mobility in individuals suffering from degenerative joint diseases such as OA. As global populations age, the incidence of these procedures continues to rise, placing increasing demands on healthcare systems. The epidemiology of hip and knee replacements reveals not only the growing burden of MSK conditions but also the critical need for improved outcomes, efficient healthcare delivery, and cost-effective care.

Advances in modelling techniques using AI and ML methods have emerged as powerful tools to transform healthcare practices, offering the potential to improve patient care, surgical planning, and predict outcomes. By analysing vast datasets, AI and ML techniques can identify patterns in patient history which contribute to more personalised and timely interventions.

The integration of EHRs has further enhanced the ability to collect and manage patient data in real time, providing clinicians with comprehensive, up-to-date information to make informed decisions. EHRs, when paired with AI methods, enable predictive models that can improve both the decision-making process and long-term patient outcomes. Furthermore, the application of graph-based AI models to EHRs opens new avenues for understanding complex relationships within patient data, helping to identify trends and predict patient outcomes.

This chapter aims to explore the intersection of prediction models, EHRs, and graph-based AI methods, within the context of hip and knee replacements.

2.2 Healthcare Background

2.2.1 MSK Conditions

The MSK system supports, stabilises, and enables movement within the human body, consisting of bones, muscles, cartilage, tendons, ligaments, and other connective tissues. MSK conditions affect these structures, leading to temporary or long-term functional impairments that can hinder daily activities. Characterised by pain, reduced mobility, and limited dexterity, these conditions significantly impact quality of life [16].

There are over 200 recognised MSK conditions [17], broadly categorised as mechanical or autoimmune-related. Mechanical conditions, such as OA, tendonitis, and joint pain, typically result from wear and tear or injury, while autoimmune disorders, including Rheumatoid Arthritis (RA), and reactive arthritis, arise from immune system dysfunction. MSK disorders pose a significant healthcare challenge, accounting for 20% of GP consultations [16, 18]. Globally, 1.71 billion people were affected by MSK conditions in 2019 [19], and by 2022, 20 million individuals in the UK were living with at least one MSK condition [16].

MSK conditions can cause severe pain, significantly reducing quality of life and independence. Many individuals struggle with feelings of being a burden on their families, and depression is common among those with severe MSK conditions [16]. As the prevalence of these conditions rises, organisations worldwide are working to address both the personal and economic burdens they impose [20].

Osteoarthritis

OA is a prevalent joint condition characterised by the gradual breakdown of cartilage, which fails to regenerate over time [21]. This deterioration is typically driven by ageing or joint injury [16]. OA is a severe joint disease that can significantly impact mobility and often leads to disability [22]. It affects various joint tissues, including bone marrow, ligaments, cartilage, and the meniscus [23].

The knee is the most commonly affected joint, with 5.4 million of the 10 million UK patients over 45 years old diagnosed with knee OA [16]. Among individuals aged 60 or older, the condition affects approximately 10% of men and 13% of women [24]. Knee OA is characterised by structural changes in osteochondral tissues, resulting in pain, stiffness, and reduced mobility.

The hip and hands are also frequently affected, with 3.2 million people in the UK diagnosed with hip OA [25]. The condition is associated with joint swelling, decreased physical activity, and increased morbidity [23, 26]. Like most MSK conditions, risk factors include sex, age, obesity, injury or joint overuse, metabolic disorders, and genetics [24, 27, 28, 29].

Symptoms can be used to define OA or alternatively radiography findings such as joint space narrowing, osteophytes or subchondral sclerosis. Dell'Isola et al. [23] have described six phenotypes of OA including: chronic pain with central mechanisms, high levels of inflammatory biomarkers, metabolic syndrome, local tissue metabolism alteration, mechanical malalignment, and minimal joint disease phenotype. The clinical classification criteria of the American College of Rheumatology provide guidelines that are commonly used for OA diagnosis [30]. The criteria requires a patient to have pain on most days in the last month, no crepitus (sounds when moving a joint such as grinding or popping) and bony enlargement, crepitus and morning stiffness for less than 30 minutes, crepitus and morning stiffness and bony enlargement. OA is likely to be diagnosed using non-radiographic information, such as joint pain severity, stiffness, restricted movement, and disability. Findings from X-rays have been shown to be ineffective towards making decisions towards joint pain management [31, 32]. However, in some cases clinicians prefer to only confirm hip OA if there is radiographic evidence of OA as clinical presentation can vary, including osteophytes or joint space narrowing [32], therefore primary care diagnosis can be more difficult. Sometimes pain location and severity do not correlate with radiographic findings [33].

OA is commonly graded using the Kellgren-Lawrence (K-L) grading system [34]. However, systems such as K-L grading can be misinterpreted due to its reliance on subjective radiographic interpretation, overlapping grading criteria, lack of consideration for clinical symptoms, limited sensitivity to early changes in OA, different X-ray angles and positioning, and variability in application across different joints or patient populations [35].

Peat et al. investigated the agreement between the GP diagnosis of OA, the patient's own diagnostic attribution, and with the American College of Rheumatology criteria. Neither GP diagnosis (actual agreement = 64%, kappa = 0.28) nor a patient's own diagnostic attribution (actual agreement = 30%, kappa = -0.39) related strongly to the clinical classification criteria, nor did they relate well to each other (actual agreement = 49%, kappa = -0.03) [31]. This could be due to the impact on a patient's life being the key determinant of the clinical importance

of the condition. There is a large gap between the first presentation and symptoms of OA to the diagnosis of OA [21]. These diagnostic inconsistencies suggest that there may be a need to improve clinical decision making procedures.

In 2019, 344 million people were living with a diagnosed case of OA, marking a 114% increase since 1990 [19]. The condition is not only physically debilitating but also has a significant mental health impact, with around 20% of individuals experiencing depression or anxiety [36]. OA is a leading cause of hospitalisation, ranking as the fourth most common reason for admission in 2009 [37]. Additionally, 25% of adults are expected to develop symptomatic OA during their lifetime [37]. As the prevalence of OA continues to rise due to aging populations, the strain on healthcare systems is expected to increase [38, 39]. Managing OA poses a major challenge for healthcare services. In 2015 alone, the UK spent £10 billion on OA treatment and care [40].

Key Outcomes from MSK Conditions

OA has a significant impact on daily life, leading to pain, fatigue, low mood, reduced independence, and mobility issues [16]. These challenges are often interconnected, for example, joint stiffness after sleep can make it difficult for individuals to get out of bed, while mobility limitations can contribute to social isolation. Pain and restricted movement are the primary factors driving disability, loss of independence, and a reduced quality of life in those with MSK conditions. MSK conditions can accumulate and lead to other health issues, resulting in long-term conditions that will eventually result in frailty [41]. Better treatment options need to be available to slow joint deterioration, and to do this not only does awareness of risk factors need considering, but better prediction tools are required to target patients before it is too late to treat them.

Between 2013 and 2014, MSK conditions cost the NHS £4.7 billion [42]. The impact extends beyond healthcare costs, significantly affecting the workforce. In 2021, 23.3 million working days were lost due to MSK conditions [43], and in 2022, 13% of Employment Support Allowance claimants in the UK had an MSK condition [16]. Individuals with arthritis are also 20% less likely to be employed compared to those without the condition [44]. Collectively, these factors create a substantial economic burden on individuals, healthcare providers such as the NHS, and, ultimately, taxpayers.

Current Management and Treatment of MSK Conditions

The progression of OA is rarely preventable at present, however, its outcomes can be substantially improved with earlier treatment and supportive interventions, including for co-morbidities. Joint pain is normally managed within primary care, where first patients are encouraged to exercise, undergo physical therapy and, where appropriate, weight loss is suggested [21]. After lifestyle changes are suggested oral Non-Steroid Anti-Inflammatory Drugs (NSAIDs) or opioids (such as codeine and co-codamol) may be prescribed and pharmacological management undertaken to reduce symptoms and relieve pain [21].

Muscle strength around the affected joints and endurance testing on patients can be carried out to determine OA severity. Disease progression is usually monitored via radiographic changes, such as joint space narrowing and osteophyte formation. Clinical trials and research studies measure OA severity using tools such as the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), however this is not currently used in clinical practice.

Patients with the highest risk of poor quality of life are referred to specialist care services. Treatment for severe and end-stage OA primarily involves total joint replacement which sometimes requires future revision surgeries [45]. Hip and knee replacement procedures are performed when conservative treatments like medications, physical therapy, or lifestyle changes no longer provide relief from pain or disability caused by joint damage. Hip and knee replacement surgeries are a common procedure, which are often elective as patients wish to reduce pain and restore mobility.

Most patients regain mobility within weeks after hip replacement surgery, though full recovery may take several months. Similarly, knee replacement recovery involves physical therapy and rehabilitation to restore joint function, with most patients resuming normal activities within a few months [46]. Both procedures have high success rates, significantly improving quality of life by reducing pain and enhancing joint functionality. Prosthetic joints are generally durable, with fewer than 12% of patients requiring revision surgery 20 years after a primary cemented hip replacement, while up to 25% of patients need revisions following uncemented hip replacements over the same period [7].

Intervention measures could be recommended earlier to at-risk patients to enable lifestyle alterations. The risk of knee replacement at time of knee OA diagnosis increases up until age 62

then decreases, whilst hip replacement at time of hip OA diagnosis decreases linearly the older the patient is [47]. Due to growing OA cases, unlicensed prescriptions are increasingly issued as clinicians seek alternative painkillers to alleviate patient suffering [48].

2.2.2 Hip and Knee Replacements

Hip and knee replacements are surgical procedures in which damaged or arthritic joints are replaced with artificial prosthetic components to relieve pain, improve function and mobility, and enhance quality of life. In both procedures, the damaged joint surfaces are removed and replaced with prostheses made of metal, plastic, or ceramic. A hip replacement (hip arthroplasty) typically involves replacing the femoral head (the ball of the thigh bone) and the acetabulum (the socket in the pelvis), while a knee replacement (knee arthroplasty) involves resurfacing or replacing parts of the femur, tibia, and sometimes the patella. These surgeries are commonly performed for severe arthritis, fractures, avascular necrosis, or joint degeneration that cannot be managed with non-surgical treatments [33, 49]. Joint replacement may involve total replacement of the joint, where the entire joint is replaced with artificial components, or partial replacement may be sufficient where only part of the joint is replaced. Surgical techniques are improving to become less invasive with better prosthetic materials, which leads to more people opting for joint replacements to improve their quality of life.

2.2.3 Epidemiology of Hip and Knee Replacement

Hip and knee replacements are most commonly performed to reduce pain and mobility problems caused by OA and other joint-related issues such as RA and trauma-related damage.

Each hip replacement procedure costs the UK NHS an average of £5,280, resulting in an annual expenditure of £438.9 million for primary hip replacements alone [38, 50]. In 2007, the average cost of major hip procedures was £7,800, while major knee procedures averaged £4,471 [17].

The incidence of hip and knee replacements has been increasing worldwide, especially in older populations, with demand for hip and knee joint replacement in the UK expected to rise by nearly 40% by 2060 [51]. In 2014, 83,125 primary hip replacements where recorded by the National Joint Registry [38], whilst in 2023 108,558 hip replacements were recorded. Similarly in 2013 82,267 primary knee replacements were recorded by the National Joint Registry [52], with 116,845 knee replacements recorded in 2023 [53]. This showed an increase of 130.6% and

142.0% for primary hip and knee replacement procedures in nine and ten years respectively.

Socioeconomic status also has an effect on the risk of hip and knee replacement. Individuals with higher IMD scores (less deprivation) are less likely to have hip replacements [6]. High-income countries are seeing the most significant increases in hip and knee replacement surgeries due to increased life expectancies and improved access to healthcare [54]. There is also an observed inequity in access to hip replacements: individuals from less deprived backgrounds are generally less likely to develop OA and, consequently, have lower overall rates of hip replacement. Yet, for those who do develop OA, they are more likely to receive hip replacement surgery compared to individuals from more deprived backgrounds, highlighting disparities in treatment accessibility. Countries with healthcare systems such as the UK's NHS are seeing higher replacement surgeries due to standards of care based on need, whereas other countries which require insurance and healthcare access have more procedure variation [54].

Knee replacement risk peaks at 60-70 years of age, and hip replacements peak at 55 [38, 47]. Women are more likely to have joint replacements, comprising 60% of hip replacement surgeries and 57% of knee replacements [38]. This could be due to factors such as higher prevalence of OA in women and differences in biomechanics [38, 55]. Whilst women are more likely to be recommended to have a hip replacement, when indications are present, women are less willing to undergo the surgery [56].

Inflammatory diseases, such as RA can lead to joint degeneration, increasing the risk of requiring a hip or knee replacement. However, patients who have RA are more likely to have complications, poorer recovery, and further pain after replacement surgery [57].

Obesity is a major risk factor for knee and hip OA, increasing the likelihood of joint degradation due to increased weight on the joints [27]. Lifestyle may also impact joint replacement risk, for example, those who participate in extreme sports or those with poor muscle strength due to insufficient exercise [58]. Trauma or fractures can result in the need for a hip or knee replacement. Older adults are at higher risk for having falls, which increases the risk of hip fractures [59].

OA is the primary indication for hip and knee replacement, accounting for 92% of hip replacements and 98% of knee replacements in 2023 [7]. The increase of hip and knee replacements is largely due to the ageing of global populations, the increasing prevalence of joint conditions

like OA, and advances in surgical techniques and prosthetic technology.

2.2.4 Clinical Pathways for Joint Replacement

There are several steps involved before hip or knee joint replacement is recommended. First, a patient's history, age, diagnoses, and current symptoms must be evaluated by a clinician, alongside imaging being required to determine joint damage extent. Patient-Reported Outcome Measure (PROMs) can be used to gauge patient pain and disability caused by the affected joint, enabling clinicians to understand their current quality of life.

If a patient has significant pain or mobility issues then conservative treatment will be recommended, including options for physical therapy, anti-inflammatory prescriptions, steroid injections, pain relief, and lifestyle modifications. If a patient responds poorly to conservative treatment and does not receive significant improvements to their quality of life, hip or knee surgeons will consult with the patient to determine suitability. Surgery decision making involves informing patients about the procedure, risks, lifestyle changes required and likely outcomes. The surgeon will decide on what approach to take to replace the joint (partial or full), alongside which type of prothesis to use.

Determining the optimal timing for joint replacement intervention is challenging because it requires balancing the patient's pain and functional limitations with the risks and benefits of surgery, considering factors like age, health, and overall quality of life. Additionally, the progression of joint degeneration can vary among individuals, making it difficult to predict when non-surgical treatments will no longer provide sufficient relief and when surgery will yield the best long-term outcomes.

Immediately after surgery the patient will be monitored and pain killers administered. The patient may be required to stay in hospital for up to a week and start physiotherapy a few days after surgery to enhance recovery. It can take months to restore joint strength, flexibility and mobility after surgery. Follow-ups with surgeons may be required to ensure the joint is functioning as expected. After hip and knee replacement recovery patients can resume normal activities, with surgeon advice on any activities that may cause joint deterioration [21, 33]. Hip replacements have been shown to last for 25 years in 58% of patients [60], whilst knee replacements last 25 years in over 70% of patients [61]. Patients may need revisions where excessive wear or dislocation has occurred.

2.2.5 Electronic Healthcare Records

When a patient visits a GP or a hospital, EHRs are used; clinicians add notes to these records depending on what issue the patient has. A clinician may write notes in the form of free-text, or they might use a particular code to denote a symptom, diagnosis or prescription.

The Format of EHRs

EHRs were initially designed for the sole purpose of assisting with medical administrative tasks, collecting huge amounts of patient data within specific medical visits [62].

EHRs are a large electronic collection of patient health data, including health events and information along with their relevant timestamps. Each patient will have an individual record in their name with temporal sequences of their clinical variables based on health visits. Procedures, test results and diagnosis of the patient are just a few of the things included within an EHR, alongside where and when the patient had their appointment and who performed the service [63]. EHRs are used commonly by clinicians to track patient medical history and enable patient risk prediction, diagnosis and prognosis. The NHS and other medical bodies continue to rapidly accrue large quantities of health data, that is yet to be fully utilised in patient care. EHRs contain information that informs disease trajectories that could be used to improve healthcare outcomes for early detection or faster diagnosis.

Applications and Limitations of using EHRs

Due to the adoption of health information technology, frontline clinicians now rely on EHR data as a critical tool for decision-making. However, data entry can be challenging, especially in fast-paced environments. For example, in the UK, NHS GPs often have less than 10 minutes per patient. This limited time leads to rushed discussions, incomplete notes, and restricted coding. GPs must speak with patients, determine their concerns, decide on actions, and record notes in the EHR system, all within a short window. This pressure is exacerbated as consultation rates rise [39, 64].

Structured EHR data, such as clinical codes, are often easier to analyse statistically compared to unstructured textual data which requires more complex processing via Natural Language Processing (NLP) techniques [65]. However, reliance on structured data poses challenges: clinical codes may be assigned incorrectly, omitted entirely, or fail to fully capture a patient's symptoms.

For example, rare diseases and health events are often under-represented in EHRs, complicating their use for machine learning models [62]. Free-text clinical notes offer richer information but require sophisticated processing for integration into predictive models.

Multiple studies highlight the limitations of EHR coding accuracy. Research on computerised coding of hip osteoarthritis (OA) diagnoses in Clinical Practice Research Datalink (CPRD), a longitudinal UK database with records from approximately 14 million primary care patients, found that while diagnostic accuracy was generally sufficient, the date of diagnosis varied significantly between GP records and CPRD records. Discrepancies ranged from 18 to 1,448 days, with errors mostly being the CPRD record being later than the actual recorded date. This finding emphasises the importance of caution when using such data for research [32].

The temporality of EHR data is another underutilised feature that could enhance predictive algorithms. Events closer in time are often more relevant to a patient's current health than those from years earlier. For instance, frequent visits in short intervals may indicate a worsening condition compared to more sporadic visits [63, 66]. Predictive models need to prioritise recent events while balancing the inclusion of historical data that might reveal patterns missed by clinicians.

ML models using EHR data show promise for improving healthcare outcomes. For instance, predicting Intensive Care Unit (ICU) admissions after surgery using preoperative data can optimise resource allocation and patient care [67]. However, the performance of these models depends on how far in advance predictions are made. A study predicting eight types of cancers found that accuracy decreased by 8.15% when extending predictions from immediate to 24 months in advance [68].

As of October 2023, 5.4% of the UKs population opted-out of their EHRs being shared beyond their own healthcare use [69]. Opt-out rates for NHS data sharing fluctuate due to public awareness, government policies, trust in data security, demographic differences, and perceptions of the benefits of data-driven healthcare research. When opt-outs increase, AI models that rely on EHRs for training and analysis may face challenges. Reduced data availability can lead to biases in predictive healthcare algorithms, affecting their accuracy and representativeness across different patient demographics [70]. This can limit the ability to generate robust insights, impacting medical research, treatment recommendations, and public health strategies.

Overall, EHRs are full of information with the potential to revolutionise healthcare. To fully realise their potential, challenges like coding inaccuracies, temporal irregularities, and privacy concerns must be addressed. Advances in synthetic data, NLP, and graph-based methods such as HORDE, which identifies missing or erroneous coding, are promising steps forward [71]. By addressing these challenges, EHRs can better support predictive modelling, enhance clinical decision-making, and improve patient outcomes.

Missing Data

As the data is not collected specifically for research, EHR records often are known for their missingness or incompleteness [65]. Missingness in EHRs refers to the absence of information within a patient's health record, potentially due to not being recorded, the patient not being asked about a symptom or condition, being recorded incorrectly, being left blank, privacy concerns, patients opting out, multiple care systems per patient, or information loss caused by technical issues (e.g. electronic system or data transfer issues). To name a few examples of data that may be missing: demographic information, vital signs (such as blood pressure, heart rate), laboratory results, or reason for visiting may be missing. That being said, now that EHRs are electronically recorded, deciphering poor handwriting or sifting through hundreds of pieces of paper to find necessary documentation is no longer necessary, most things needed can now be found at the click of a few buttons. Additionally, EHR data is a more affordable option compared to specifically collecting data [65].

When analysing healthcare data it is important to address missingness carefully and appropriately to prevent biasing or inaccuracy of the results. Missing data can be dealt with by using statistical methods such as imputation, which involves estimating the missing values. Missingness may be represented in a model by including an extra variable that indicates the presence of missing values. It is also vital to consider that a condition might not be recorded simply because someone does not have that condition, or they might not have the condition documented [65].

There are four mechanisms of missingness (missing completely at random, missing at random, missing not at random, missing by design). Missing completely at random means that data is independent from other information and the missing data would not bias the analysis. Missing at random means that the missingness only depends on observed data and not the missing data, this assumes that the missing data does not influence the analysis after considering the observed

variables. Missing not at random means that the missing values are different to the observed data, as such the missing data could contain useful information that could affect the analysis. The final type of missingness mechanism, missing by design, is where data is purposefully missed during data processing, this informs what data were intentionally collected. Many imputation methods idealise that the data is missing at random or missing completely at random to make computation simpler [72].

Missingness can be simply dealt with by removing the variables which have missing data, or removing patients (complete case analysis) with missing data (whichever results with the least amount of information lost). However, if excluding patients is necessary there should only be a small number of patients that have the variable missing and it must be considered carefully whether the missing data is random, if it is not at random care needs to be taken to ensure that this does not introduce bias into the model. A variable should be considered for rejection if it has many missing values or the variable is missing not at random; otherwise, the reason for missingness should be thoroughly investigated, calculating uncertainty and analysing sensitivity to see the effect on the final imputation and to reduce bias [72].

Missingness in EHRs can be be informative, particularly when incorporating into prediction models. The lack of a variable may be informative of a patients health status, for example a patient with severe diabetes may miss more routine check ups due to being hospitalised, whereas someone with better managed diabetes may have more regular visits and therefore more complete records. As such, frequency of recording may also be an informative form of missingness. Cholesterol is not normally recorded by a GP, therefore if a patient has multiple recordings for cholesterol it is likely due to the GP suspecting an underlying issue which may inform that the patient might have poorer cardiovascular outcomes [73]. The likelihood of missing data may correspond to disease severity, such that patients with poorer health likely have more conditions, whilst a healthy patient will utilise health services less [65].

Imputing missing clinical codes in EHR data presents a challenge due to the irregular time intervals between recorded entries. Unlike structured numerical data, the absence of a clinical code does not necessarily indicate missingness but rather a lack of documentation or clinical relevance at the time of recording. To accurately infer missing codes, NLP techniques would be required to extract relevant clinical information from unstructured EHR text, alternatively a model could be trained to learn clinical code trajectories and impute codes where commonly

there would exist a code, however this could lead to bias in the model due to trajectory manipulation. Without such an approach, it is impossible to determine which clinical events or conditions might be absent from the structured record, as with clinical codes alone it is impossible for someone to know what has not been documented.

2.3 Technical Background

2.3.1 Knowledge-based Risk Prediction Models

Risk prediction models can enable outcomes to be detected or predicted earlier than a human may detect a pattern. Early prediction is especially important in healthcare applications to allow timely action to be taken. Additionally, risk prediction modelling tools may enhance resource allocation, improve surgical outcomes, improve patient satisfaction by improving life quality, free up clinical time to focus on being empathetic and helping patients.

Associative rules can be used for predictive models. For example, Folino et al., 2010 built a comorbidity network and defined the risk of specific diseases occurring in patients using their patient record [74]. Confidence rules are simply used to determine the probability of someone getting a specific disease once their historical health data using population prevalence has been provided [74].

Other knowledge-based risk prediction models may use rule-based systems following 'if' and 'then' type statements in a chain to arrive at a risk score, logical flows including data may be used, alternatively probabilistic reasoning can be carried out by representing data as networks or graphically [75]. Whilst these methods can be useful for flexibility and simplicity, with the rising amounts of healthcare data being generated, manually creating knowledge-based risk prediction models can be highly challenging to develop.

2.3.2 Machine Learning and Artificial Intelligence

Compared to traditional risk calculators used in the health setting, ML techniques offer the ability to utilise significantly larger numbers of variables for predicting clinical outcomes [67]. These variables are typically already present in EHRs, potentially enabling automatic clinical decision support to be generated for clinicians.

ML is a subset of AI. AI broadly aims to simulate intelligence and has deeper and more human-

mimicking applications than ML [76]. For example, AI is the foundation of Chatbots, virtual assistants, and self-driving cars, whilst ML is typically used for things such as social media recommendations or filtering fraudulent transactions.

ML can be defined as models that can learn from data automatically. ML involves using computer algorithms which find patterns from inputted data, to provide probabilities and predictions of an outcome. Training of ML models is done by optimising parameters to minimise the error between the predicted and actual or expected output.

A systematic review of 145 low risk of bias comparisons of logistic regression versus ML in healthcare applications, demonstrated that the Area Under the Curve (AUC) difference was on average 0.00 [77]. Notably, papers with high risk of bias reporting in 145 comparisons had a AUC difference of 0.34 in favour of ML. Nonetheless, some studies claim that ML algorithms have better performance over tradition statistical modelling (such as logistic regression) when a huge number of predictors are utilised or if the ML model is more complex [78]. ML models can be more data hungry and require larger quantities of data than regression models to obtain similar performances. But there is little evidence supporting the claim that ML models have better discriminative performance over regression models. Additionally, linear and logistic regression models are less likely to overfit than neural networks due to regression models having limited hyperparameters [79].

ML and AI models typically have two types of parameters, learnable parameters and hyperparameters. Learnable parameters adapted internally within models iteratively, using the training (and validation if applicable) set(s). For example, with neural networks the learnable parameter would be the weight of each neuron [80]. Hyperparameters, for example the number of layers in a network, learning rate, or batch size, are adjusted externally from the model and can changed by the coder manually or iterated through using grid search techniques. Once both the learnable parameters and hyperparameters have been optimised to achieve a model with the desired performance, the test set can be evaluated with these parameters to assess the model.

Traditional Machine Learning Models

Regression models, such as as linear and logistic regression models are used to predict values, probabilities, or binary outcomes. Linear regression is a supervised learning algorithm that predicts a continuous output by modelling the relationship between independent variables and

a dependent variable as a straight line. Logistic regression is used for binary classification tasks, modelling the probability of an outcome as a sigmoid-shaped curve to distinguish between two classes [81].

Risk prediction models may be formed from statistical methods to produce risk scores based on a range of predictor variables. For example, Chan et al., 2018 developed the Combined Assessment of Risk Encountered in Surgery (CARES) which is a surgical risk calculator that uses 9 variables to predict whether a patient will need to stay in the ICU for more than 24 hours and their 30-day postoperative mortality risk [82]. This uses multivariate logistic regression to obtain odd ratios which are then provided as rank scores.

Logistic regression approaches enable calculation of the odds ratios and standardised coefficients for independent variables [79]. Which, after adjustment, enables a degree of interpretability of which variables have more effect on the prediction.

Decision trees are a supervised learning method that splits data into branches based on decision rules derived from features, creating a tree-like structure to classify or predict outcomes. Random Forests (RFs) improve on decision trees by combining the predictions of multiple trees, reducing overfitting and increasing accuracy through ensemble learning [83].

Support Vector Machines (SVMs) are supervised learning algorithms used for classification and regression tasks, where they find the hyperplane that best separates data points into distinct classes by maximising the margin between them. SVMs are particularly effective in high-dimensional spaces and can handle both linear and non-linear boundaries [84].

K-means and other clustering methods are used as unsupervised learning to group similar patterns and features into groups [85]. For example, to cluster for multimorbidity to see which morbidities commonly occur together [86]. Clustering tasks have the the potential to reduce the number of parameters required for the model by combining commonly occurring conditions. Unsupervised methods are also highly sought after as manual labelling of data can be time consuming and resource heavy [85].

Deep Learning Models

Neural networks are typically made up of layers of neurons interconnected to each other, which takes an input and passes it through the layers to make a prediction. Artificial Neural Networks

(ANNs) were one of the first deep learning methods. ANNs are made up of at least three interconnected layers of neurons, the input, hidden layers (1+ layer(s)) and the output layers. ANNs are an extension from linear regression and can be used to model input and output relationships which are non-linear [87]. Deep learning typically involves automatically extracting features and learning patterns in data using multiple layers of neurons (neural networks) which transform the data non-linearly.

Recurrent Neural Networks (RNNs) are used to for time series data and for natural language processing. RNNs have inbuilt repeated connections which have time distributed within its hidden states, so that previous information can be retained and passed forward whilst temporal dependence can be learnt [88], similarly to how words in a sentence are used together rather than individual words to give meaning. As RNN models train using back propagation, vanishing or exploding gradients can occur when the contextual information inputted into the RNN is lengthy [89]. Additionally, the time that elapses between events is ignored with RNNs.

Long Short-Term Memorys (LSTMs) are an extension of RNNs, with LSTMs able to pick up on longer-term patterns. LSTMs can learn long-term dependencies from sequenced data and take more previous data into account when making a prediction, due to retaining information from lengthy periods of time. This is done using the same control-flow as RNNs, with the exception of the LSTMs having memory blocks/LSTM cells which chooses random time intervals to retain information [88]. Three gate units are used within the cell to either update (via the Input gate), keep/remove (via the Forget gate) information, or output information via the Output gate. These cells contain a sigmoid activation function σ, which outputs a value (between zero and one) to determine how much is passed through the cell, where one means all information can be passed through and zero would mean no information should be let through.

CNNs are commonly used for data with spatial structures such as images, for object detection and image classification. One-dimensional CNNs can be used for processing time-series data, for example language. Two-dimensional CNNs are the most common and are typically applied to images. Three-dimensional CNNs can also be used for movement analysis and videos. CNNs contain hidden layers which perform convolution operations which uses filters (also known as kernels) to detect patterns. The filters can be used during the convolutions to extract important and common characteristics amongst the sequential input data [87]. Deeper CNNs are able to detect more complex features than shallow CNNs. More details on CNNs will be provided in

Chapter 4.

Transformers are commonly used for NLP tasks, BERT and GPT are just two of the popular language generation models that are being used widely. Transformers enable processing of images with very little effort, the input image is simply split into patches and fed in sequentially based on the pixel patches. Transformers utilise global self-attention networks to give specific weightings to different parts of the input data. However, the global attention design of Transformers means that larger inputs can have performance problems. CNNs have been shown to perform superiorly to Transformers in image prediction tasks [90], suggesting that self-attention is not the factor that leads to significant performance increase compared to other models.

Other Models

Bayesian networks are graphical models that represent probabilistic relationships among a set of variables using Directed Acyclic Graphs (DAGs), where nodes are variables and edges indicate dependencies. They are commonly used for reasoning under uncertainty, enabling tasks prediction and decision-making in complex scenarios [91].

Reinforcement learning uses agents to teach programmes optimal actions by reward and penalty giving mechanisms, for example to learn best decisions to win a video game or to optimise robot movements and automations [92].

Models such as Graph Neural Networks (GNNs) are useful for processing data which is represented as graphs. More information on graphs and their uses is presented in Section 2.3.7.

2.3.3 Predicting MSK Conditions and Outcomes

Predicting hip or knee replacement in advance could potentially reduce the time patients spends in pain whilst on a surgical waiting list. Early detection of MSK conditions provides many advantages to patient care along care pathways, motivating the need for models to predict MSK conditions and outcomes as early as possible to improve treatment plans [1].

The development of an individualised risk prediction model for long-term OA progression, may be used to develop bespoke intervention plans for individual patients, with the potential to delay OA if exact aetiologies and similar clinical presentations are discovered. Prediction models incorporating predictors such as risk factors, symptoms and image features to discover patients with higher chances of developing late stage OA, could be beneficial to reduce population pain

by OA prognosis. Treatment plans could be put in place for at-risk patients in cases where OA progression can be delayed [80].

Clinical radiology can be enhanced using predictive modelling to improve radiologist efficiency and precision by optimising imaging protocols to give faster and automatic reporting [80]. OA classification has been performed using deep learning models on clinical images, such as radiographs and Magnetic Resonance Imaging (MRI) [22, 80]. Kellgren-Lawrence (KL) knee radiographic scores can be calculated from radiograph images to indicate severity. MRI can be used to find menisci tears, cartilage damage, and show changes to cartilage biochemistry [22]. Approaches using radiographic images and clinically recorded factors such as family history, prescriptions, general health and joint pain have also been performed to assess OA severity [93, 94]. Tools for OA severity prediction which include MRI have been shown to obtain higher accuracy (AUROC=0.72) compared to tools which only contain clinical notes and demographics (AUROC=0.6) [22]. Using deep learning models on clinical images has the advantage of acting as a clinical decision assistance tool, alongside finding abnormal findings faster, leading to increased throughput of images analysed in a given day. These deep learning models may perform classification of an image to give a severity grade, or may even have the ability to segment an image to find an area of interest, for example regions where the cartilage has severely degraded [80].

Risk prediction of occurrence of knee OA one year in advance using EHR non-image data alone, has been done achieving AUC scores of 97%, from three years worth of sequential diagnoses, prescriptions, sex and age [87]. This clinical data was inputted into models containing CNNs and ANNs, achieving AUC scores of 0.97 when using this method. Ningrum et al., 2021 showed discriminative features from their study, suggesting that age and sex were not ranked as highly important features compared to prescriptions (such as cough suppressant, expectorants and antacids) and diseases (eye disorders, connective and MSK diseases, metabolic and immunity disorders).

AI models can be utilised to build patient-specific tools to detect MSK conditions earlier to enable preventive treatment which can reduce the number of people with disabilities and pain. A variety of reviews of prediction models for MSK clinical applications have been undertaken, covering prediction of outcomes such as of knee OA [80, 95], hip OA [95], physiotherapy recommendations [96], and other MSK joint related pains [95]. Current models have limited success in

matching predictions to outcomes, with model validation performance requiring improvement before patients see treatment and quality of life enhancements [96]. Other models predicting hip and knee replacement are covered in more detail within Chapter 6.

2.3.4 Process Mining

Process mining techniques analyse sequences of events over time to create process models, which reveal relationships and patterns in workflows. Event logs, typically used for these models, contain descriptions of activities, timestamps, and case identifiers, with optional metadata such as event types or resources [97]. Process mining is valuable in healthcare, where it can uncover disease trajectories, optimise care pathways, and improve decision-making [98]. For example, process mining has been used to understand frailty progression [99], optimise multidisciplinary collaboration for diabetes care [100], and analyse oncology treatment pathways [101].

However, process mining techniques like heuristic and fuzzy mining often generalise patient data into single models, making them less suitable for individualised outcome predictions [100]. While these methods provide insights into group trends, this research focuses on individualised predictions using graphs derived directly from EHRs.

DAGs offer a complementary approach, sharing structural similarities with process models but focusing on patient-specific temporal relationships. Unlike process mining, which often uses unsupervised learning for discovering trends, this project employs supervised learning to predict clinical outcomes at the individual level. By converting EHRs into graph representations of clinical codes and inter-event times, the method described in this thesis captures detailed patient histories without requiring generalised process models.

2.3.5 Explainability

Explainability in the context of AI is how the decision making processes of a model can be provided to the user of a model. Such that the user understands what model predictors or combination of predictors led to the model prediction outcome. For example, predictors might be assigned scalar values showing their influence on prediction or heatmaps of medical images may be used to show regions or features the model finds most impactful for prediction. Deep learning models often have many hidden layers and complex processes so are ‘black-boxes’ usually with little to no explainability, whilst ML or traditional statistical methods often offer

more interpretable results.

Dissimilar to other uses of AI, the interpretability of AI for healthcare comes in precedence to the model's predictive accuracy. This is because AI in the healthcare setting serves as clinician decision making assistance tools, rather than tools free of human input.

There are methods that can be applied to AI models, either during training or post-training to provide explainability. Two simple and commonly used techniques include Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME is a method which shows the features to that contribute positively or negatively to a prediction outcome [102]. LIME has been used alongside RNNs and attention mechanisms on EHR data to predict heart failure, looking at the patient visits most likely to be influential to the model prediction [102]. SHAP is commonly used with tabular data to show global clinical features which correspond to a positive or negative outcome [103]. Models such as RFs, SVMs and logistic regression models can be used with SHAP to get increased interpretability, for example to predict COVID-19 [104].

Zhang et al. 2019 designed an attention-based time-aware LSTM Network (ATTAIN) model for disease progression modelling. Zhang et al. adjust the memory of the LSTM depending on the number of prior events, decaying the weights as they become further from the present as they are likely to be less important than more recent events. The weights are used to show prior event importance relative to the current event to determine the worsening condition [66]. The ATTAIN model was used on EHR data to predict septic shock, alongside demonstrating the interpretability of their model via flagging critical event time abnormalities [66].

Chapter 7 delves into explainability methods further, including those that have been used with EHRs and graphs for health outcome predictions.

2.3.6 Artificial Intelligence for EHRs

In this thesis, EHR data is focused on as the data for health outcome prediction as EHRs offer several advantages over other types of data. EHRs provide a rich and comprehensive source of clinical information, including diagnoses, lab results, medications, vital signs, and treatment histories, offering a holistic view of a patient's health. This is further enhanced by the longitudinal nature of EHRs, which track patient data over time, allowing AI models to identify patterns and predict disease progression or the need for early intervention. EHRs

are already integrated into clinical workflows, making it easier to deploy AI-driven predictions directly in real-time decision support systems.

Artificial Intelligence for EHRs in the NHS

The NHS has been exploring and implementing AI technologies for EHRs, particularly through initiatives like the NHS AI Lab and the AI in Health and Care Award. However, despite using AI in radiology, diabetes management, long-term hospital prediction, amongst other outcomes, the use of AI in EHRs within the NHS is still limited, with a strong emphasis on ethical considerations, explainability, and alignment with regulatory standards like the NHS Digital Technology Assessment Criteria (DTAC). Projects are also focused on interoperability and improving patient access to their own health data through tools like the NHS App, which could eventually integrate more AI-powered functionalities to aid self-care and disease management.

One of the primary challenges in applying AI to healthcare is the lack of robust validation and limited model reliability. Many models exhibit high Risk of Bias (RoB), resulting in poor performance when applied in real-world contexts and demonstrating minimal generalisability. For instance, a systematic review and appraisal of 66 predictive models for COVID-19 revealed significant concerns. The analysis concluded that implementing these models in practice would likely lead to patient harm due to their insufficient reliability and validation [105].

Models using EHRs

AI enables us to build patient-specific models, aiming to prescribe treatment or recommendations early on in someone's disease trajectory, leading to reductions in patient pain, disability, and preventable diseases. AI techniques do this by locating patterns and associations within historical healthcare data, to enable future disease diagnosis or health outcome prognosis.

EHRs have been utilised for various predictive tasks, popularly used as predictors within models with CNNs, RNNs, GNNs, transformers, logistic regression, and Cox regression components [106, 107, 108, 109]. RFs, adaptive boosting, gradient boosting and SVMs have been compared to two risk stratification models (CARE and ASA-PS) to predict 30-day mortality and ICU stay using routinely collected data from preoperative anaesthesia assessment visits [67]. Gradient boosting performed the best for both mortality and ICU prediction with an AUPRC of 0.23 and 0.38 and an F1 score of 0.28 and 0.36 respectively, improving sensitivity by over 50%.

RNNs have been used with lists of EHR clinical codes, as RNNs are designed to learn from sequences of events. Using basic RNNs models for EHR prediction means that the time elapsed between events is ignored as these models assume uniform a distribution of time, despite potentially being clinically significant. Additionally, RNNs are less interpretable than logistic regression and other traditional models despite having higher accuracy, therefore require the addition of explainability methods such as attention mechanisms [66, 110].

The REverse Time AttentIoN (RETAIN) model has two levels of neural attention to locate clinical variables that have high contribution to the predictive outcome [110]. This model has been used to predict heart failure, and uses the assumption that more recent clinical visits are more relevant for health prediction and would therefore have higher attention influence. The variables and visits were given higher attention values if they were more influential, which is calculated based on how the input changes the probability of a specific label being predicted. The authors also showed that using timestamps (time between visits or time since their first visit) within their model resulted in a minor improvement to predictive performance [110].

Wu et al. [111] demonstrated the efficiency of their model (Order-Aware Medical SeQuence Learning (OA-MedSQL)) for predicting treatment initiation using EHR data. The model uses customised LSTM with built-in temporal sequence learning. The medical events in this model are ranked into 32 importance levels. Higher ranked neurons (more important events) are updated less frequently and stored for longer duration than lower ranked neurons which represent the importance of medical events [111]. Wu et al. [111] demonstrated that OA-MedSQL (and other sequential models such as baseline LSTMs and RETAIN [110]) significantly outperformed non-sequential models for treatment initiation prediction in this domain by over 7% (AUPRC).

The Electronic Frailty Index (eFI) model is used in UK healthcare practices to measure frailty using routinely collected EHR data [112]. The accumulation of 36 health conditions are used to calculate these frailty measures, which are then used to guide patient care. The original eFI model uses primary care EHR data. This model is used for risk stratification of patients, alongside helping clinicians create individualised care plans. This model has been adapted to disease-specific applications to assess tolerance to treatments or surgeries [113] and in community care settings for community service resource requirements [114].

The STRATIFY-Falls tool, is another tool used for to improve care and management of elderly patients [115]. It uses EHR data on things such as falls, diagnoses, and prescriptions to get

patient fall risk factors which are used to create a risk score, with a high or low risk criteria of the patient falling in the next 1-10 years. High-risk patients are flagged to clinicians as needing intervention, and then risk is reassessed using updated EHR data to help refine care plans.

Sequential data mining has been used to determine temporal patterns as clusters of medical events [116, 117], however this is an inefficient method for EHRs with long event histories [111].

A common problem with using EHRs as predictive variables in AI models is that a patients medical history may be extremely lengthy, containing thousands of medical events. If the amount of medical event combinations increases exponentially, then a problem known as *phenotype explosion* may occur, which means excessive dimensions on the data may prevent the model from fitting [118]. If these sequences of medical events are shortened to account for phenotype explosion, considering sections of a patients medical record may lead to issues such as inadequate representation of the EHR as a whole.

Integration of EHR models into clinical practice is difficult, but necessary to avoid wasting technology. Interoperability is a particular challenge on healthcare due to heterogeneity in clinical systems and patient data models. More evidence from research needs to be generated to show sufficient model performance to improve clinical decision making. Additionally, deployment of individual models are complex to manage especially when they have a narrow and specific task. Lots of AI models would be needed to cover different needs. These AI models would possibly integrate and have high technological cost. Ainsworth and Buchan highlight the potential of models using EHR data to improve health systems by enabling learning health systems that adapt to new data for better predictions and treatment management [119]. They propose real-time, personalised care models that support clinical decision-making by considering individual patient differences. EHR-based models can enhance population health management through risk stratification, preventative care, disease management, and addressing health disparities. To be effective, these models must be robust, generalisable, and capable of providing timely, clinically relevant insights.

2.3.7 Graphs

Definition 2.1. ***(Graph)*** *A* ***graph****, also referred to as a* ***network****, is defined as* $G = (V, E)$ *[120], where:*

- V *is the set of* ***nodes*** *(or* ***vertices****) in the graph, represented as* $V = \{v_1, v_2, \ldots, v_n\}$,

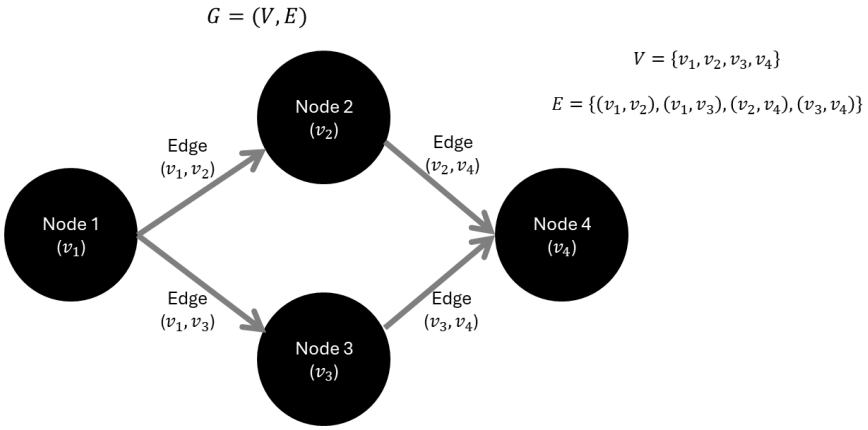


Figure 2.1: Visualisation of graphs (nodes and edges) with definitions.

where each v_i is a node.

- *E is the set of **edges** that connect pairs of nodes, represented as $E \subseteq \{\{u, v\} \mid u, v \in V, u \neq v\}$.*

For directed graphs, the edge set is represented as $E \subseteq \{(u, v) \mid u, v \in V, u \neq v\}$, where (u, v) denotes a directed edge from node u to node v.

Formally, a graph can be visualised (such as in Figure 2.1) as a structure where:

$$G = (V, E), \quad V = \{v_1, v_2, \ldots, v_n\}, \quad E \subseteq \{\{u, v\} \mid u, v \in V\}.$$

Graphs are useful ways to demonstrate networks and interactions between entities. Social media companies use graphs to represent connections between users, which they use to make friend suggestions or advertise products [121]. Online marketplaces use the interactions between customer purchases to make suggestions to other customers purchasing similar items. Veselkov et al. [122] use AI (Support Vector Machine and Maximum Margin Criterion) alongside drug-gene connections and compound-gene interactions with gene-gene connection networks to predict which drugs/foods can be used for treatment or prevention of cancer. The nodes represent the gene-encoded proteins and the edges show biological interactions. This model was also used to show the likeness between food molecules and anti-cancer drugs using hierarchical classification.

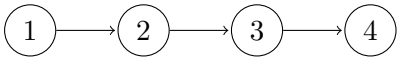
A DAG is a graph which does not contain a closed loop, each edge will originate from one node to another and the connection of these edges between nodes makes up the network structure

and flow of events. Graphs are usually visualised with a series of circles (the nodes) representing the events which are connected to each other with lines or arcs (the edges) (see Figure 2.1, to show flow of information or progression through events. This makes representing activities that happen temporally easy to conceptualise and model. If a node or edge has labels these are called attributes, which are usually provided as vectors or additional matrices. Regular graphs contain edges that connect only two nodes.

Typically, adjacency matrices are used for mathematical calculations involving the graph structures. These are usually 2D and give information about node pairs, such that rows are the start nodes and the columns are the nodes the start nodes are linked to. The share of a typical adjacency matrix is $n \times n$ where n is the number of unique nodes. For example if an adjacency matrix is:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The corresponding directed graph would be:



Classification can be in the form of node, edge or graph classification, alternatively clustering of graphs can be carried out for unsupervised learning.

Graph Neural Networks

GNNs learn graph, node and edge representations via information propagation which shares weighted node attributes between neighbouring nodes via connected edges [123]. This is used for node classification, link prediction or graph-level predictive tasks. GNNs are able to model irregular and complex graph data [62]. Graph classification predicts an outcome based on the entirety of a graph, considering the structure and attributes of the graph as a whole. Whereas edge-level prediction may predict missing or future links between nodes, likewise node-level prediction can be used for nodes.

Applying CNNs to graph-structured data may also be referred to as GNNs or Graph Convo-

lutional Networks (GCNs). GNNs iteratively aggregate information about neighbouring nodes (usually 1-hop of neighbouring nodes) to update each node representation. A GNN typically has K hidden layers, a feature vector is formed at each k-th layer, every node is aggregated and updates based on 1-hop worth of neighbouring nodes information. Within the k-th layer there are two trainable functions (aggregate and update) which could simply be summation, average or maximum. After iterating across all of the hidden (k) layers, pooling of the node representation can be achieved (which may involve averaging or more complex pooling functions) to capture structures within the graph more effectively [123]. Whilst this is a useful to aggregate information between nearby nodes, this does not fully take into account the underlying structure of the graph.

Knowledge Graphs

Knowledge graphs are powerful tools for illustrating relationships between entities, highlighting how objects interact or share similarities. Knowledge-guided methods use these graphs to address challenges arising from data insufficiency, particularly in domains like healthcare, where high-quality labelled data may be limited [124].

The GRaph-based Attention Model (GRAM) integrates medical ontologies into a knowledge graph, using RNNs and graph-based attention mechanisms to predict clinical outcomes. By embedding medical knowledge, GRAM reduces dependence on the frequency of clinical codes, ensuring robust predictions even for rare medical events [125].

Knowledge Attention Model for medical Event prediction (KAME) extends the use of knowledge graphs, employing RNN-based attention mechanisms for diagnosis prediction. Proposed by Ma et al. [126], this model efficiently combines EHR data and domain-specific medical ontologies to enhance predictive accuracy.

Building upon GRAM and KAME, the Graph Neural Network with Dynamic Propagation (GNDP) framework was introduced by [62]. GNDP enriches raw EHR data by integrating sequential medical events with medical ontology knowledge. The graph structure captures both spatial and temporal patterns, significantly improving the prediction of the next medical code during a clinical visit.

GNDP represents EHRs as hierarchical graphs where the child nodes represent sequential medical events, with vectors denoting when and in what order these events occurred. And parent

notes are derived from the medical ontology, feeding domain knowledge into the event nodes via undirected edges that denote relationships. This structure allows merging of ontological knowledge with event sequences. However, while GNDP models the sequence of events, it does not explicitly account for temporal gaps or precise numerical timestamps between events, potentially limiting its capacity to reflect detailed temporal dynamics.

Temporal Graphs

A temporal graph is a graph that incorporates time properties, for example it could evolve over time or the edges and/or nodes may have timestamps or time intervals. Many systems have time-dependent aspects, and to represent them dynamically temporal graphs are required [127]. Including temporality in graphs is beneficial as it captures node associations and when these relationships occur, increasing the detail on the domain information included to give fuller pictures [128, 129]. However, this data richness comes with issues such as scalability due to increased dimensionalities, complexity and data sparsity issues as often graphs are not fully-connected and may be incomplete [127, 129].

Temporal graphs are being used in many different applications such as on social networks to gather crime information [130], performing skeleton-based action recognition tasks to predict human actions [131], and modelling the spread of population diseases such as COVID [132].

Temporal graphs may be Static, Discrete or Temporal [133]. A Static Graph represents a set of nodes and edges with optional features, where edges are directed and features remain constant. A Discrete Time Temporal Graph restricts a temporal graph to predefined, equally spaced timestamps, ensuring all events align with fixed time steps. A Temporal Graph extends a static graph by associating nodes and edges with time intervals and time-dependent features, allowing for dynamic changes over time.

Alongside node, edge and graph classification, temporal graphs can be used for event time and temporal link prediction [133].

Using Graphs with EHRs

Graphs or network-based models can be used to show associations between different types of EHR data, with the graph edges showing the connections or relationships between these data types. EHRs may be represented as individual graphs for each patient, for each visit, or they

may be used to represent a variety of patient pathways in one graph (like a knowledge graph).

In 2019, Schrodt et al., performed a literature review of graph-representations of patient data finding eleven relevant papers, ten papers included EHR data. Five papers allocated nodes as medications and diagnoses, whilst the other six used laboratory result nodes. Edge representations most frequently included temporal relations, followed by causal relations, spatial relations, anatomic-functional relations, taxonomical relation, status, and date. Only two of these papers processed the EHR representations, for example using ML applications [134].

In the next chapter (Chapter 3), a systematic literature review is carried out on uses of graph representations for individual patient EHRs for prediction of health outcomes and diagnosis.

Researchers, developers, and implementers of complex clinical prediction models must carefully consider missingness, measurement error, and informative observation patterns in their training and test data. Causal graphs provide a structured way to reason about these issues by mapping relationships between variables based on existing domain knowledge. Causal graphs can help determine missing data mechanisms. By incorporating causal diagrams, we can visualise how errors propagate through the model and identify strategies to mitigate their impact. Informative observation patterns, that may only be present in the training or testing datasets, can introduce biases that affect generalisability. Causal graphs can be used to assess whether observed associations reflect true causal relationships or are artifacts of data collection processes.

Causal inference aims to uncover cause-and-effect relationships, whereas prediction modelling focuses on optimising accuracy. Domain experts provide crucial insights into clinical workflows, patient behaviours, and systemic biases that may not be explicitly captured in datasets. By integrating expert knowledge into causal graph construction, researchers can refine assumptions about missingness mechanisms, measurement reliability, and selection biases. This collaborative approach ensures that models are not only statistically sound but also clinically meaningful, improving their applicability in real-world decision-making. Recent studies have explored the use of directed acyclic graphs (DAGs) to represent measurement error and information bias in epidemiological research, highlighting their potential to enhance transparency and robustness in clinical prediction models [135].

2.4 Conclusion

MSK conditions, including OA, are a growing burden on healthcare systems, with hip and knee replacements being essential interventions to alleviate pain and restore mobility. The increasing demand for these procedures underscores the need for improved clinical decision-making, more efficient healthcare delivery, and cost-effective resource allocation. EHRs contain vast amounts of structured and unstructured clinical data, providing an opportunity for AI models to enhance prediction and triage for joint replacement surgery.

Graph-based AI models offer a promising approach to representing the complex, longitudinal relationships within EHR data, yet their application in predicting hip and knee replacements remains under explored. Existing predictive models often rely on structured variables such as OA diagnosis, despite a well-documented gap between symptom onset and formal diagnosis [21]. This delay highlights the need for models capable of identifying at-risk patients earlier, without requiring a confirmed OA diagnosis. Furthermore, many models neglect the irregular timing of healthcare visits, limiting their ability to capture the progression of MSK conditions over time.

While deep learning models can incorporate a broader range of predictors than traditional methods like logistic regression, their clinical utility remains limited due to issues of interpretability and explainability. Additionally, most AI models for hip and knee replacement prediction rely on imaging data, which is unavailable in primary care settings, making them unsuitable for early-stage risk stratification. Given that GPs have limited time per patient, AI-driven decision support tools could help streamline risk assessment, allowing clinicians to focus on patient care rather than data processing.

Despite the potential of AI to optimise triage, there is currently a lack of research integrating graph-based methods with EHR data for this purpose. Moreover, systematic reviews evaluating the use of EHR data with graph-based AI techniques are needed to better understand their effectiveness. To advance clinical decision-making and reduce the economic burden of MSK conditions, future research should address these gaps by developing interpretable, temporally-aware AI models that augment traditional OA diagnosis and severity (including image-derived) information with earlier-recorded and more wide-ranging information predictive of OA and its outcomes. Addressing these challenges could enable earlier and more accurate identification of patients in need of joint replacement, improving patient outcomes and healthcare efficiency.

Chapter 3

Systematic Literature Review

3.1 Introduction

Improvements in medical advances with increasing complexities of patient treatment pathways and ageing demographics have increased pressure on healthcare services. Implementing predictive algorithms into healthcare settings can reduce cognitive burdens for clinicians whilst reducing patient wait time for care [1].

EHRs are used within clinical practice to document and store patient data during clinical encounters. EHRs contain data such as health events, symptoms, laboratory investigations, and diagnoses [136]. Most clinical prediction models use summary data, such as EHR codes, which alone loses the inherent structure and temporality of the data.

Graph theory utilises network structures and uses mathematics to observe patterns and structures within data. In discrete mathematics, a graph $G = (V, E)$ is defined as a series of nodes V connected via edges E to represent relationships between nodes [137, 138].

Graphs can be used to model EHR data to maintain structural and technical features such as temporality and comorbidities, which can then be used to predict patient outcomes with ML. Earlier predictions of health outcomes may allow preventive interventions to be carried out (e.g., physiotherapy, medication), which can reduce the impact of healthcare utilisation, lessen patient suffering, or prevent conditions from worsening.

Graph representations for EHRs are becoming increasingly popular. Using graph theory, EHRs can be represented graphically to exploit the relational dependencies of the multiple informa-

tion formats to improve ML prediction models. Social network analysis methods can be used to find disease progression by finding similarities between patients and their outcome trajectories [139, 140]. Patients can be clustered based on graphical EHR representations to make diagnoses [118]. Increasingly deep learning methods, such as CNNs and GNNs, are being used to find important features and patterns in an individual patient EHR to predict patient prognoses [141, 142]. These methods could be visualised for model explainability to help clinicians understand why predictions are being made and optimise treatment plans. Understanding the factors that shape key decisions is essential for improving predictive models and ensuring their relevance to real-world clinical practice. By examining past choices, we can enhance transparency, refine the selection of predictive features, and identify potential biases or errors that may influence outcomes. This process allows for greater alignment between data-driven insights and medical judgment, making predictions more interpretable and reliable. Additionally, analysing decision patterns helps improve consistency, optimise feature selection, and strengthen the overall fairness and effectiveness of risk assessment frameworks in healthcare.

This systematic literature review follows the PRISMA guidelines for comprehensive literature search and selection. It also uses the PROBAST criteria to assess the RoB and quality of papers to investigate the utilisation of graphs in healthcare for patient-level EHR representations and health predictions. The research question guiding this review is:

How are graphs being used on EHRs to predict diagnosis and health outcomes?

To address this broader question, the following sub-questions are asked:

a) What graph approaches are researchers taking to predict these health outcomes?

b) How do these approaches compare to other ML, AI, and statistical models?

c) How are nodes and edges being utilised to perform these tasks?

d) How do these graph approaches compare to each other?

Outcomes from these studies are highly heterogeneous making a meta-analysis inappropriate. Instead, results are presented as a narrative synthesis, comparison, and discussion of studies.

This chapter seeks to explore the current literature using networks for prognostic prediction of health outcomes and diagnostic prediction of health conditions within EHRs, to determine what gaps in the literature exist and how efforts can be extended to improve clinical utility.

3.2 Related Work

At the time of this review, the review by Schrodt et al. on graph representations of patient data is the only systematic literature review paper identified that focuses on graph representations of patient data [134]. Their review of 11 articles examines how graphs were used to represent EHRs of individual patients.

There were five systematic reviews focusing on ML based prediction tasks using EHRs as input data. These did not focus solely on graphical representations. Two papers explored deep learning models using EHR, but neither retrieved any graph-based models [10, 143].

Si et al. review on deep representation learning of EHR identified 49 papers [144]. They discussed graph-based patient representation, models such as GNNs, and highlighted various works (n = 8). Three references in their paper also match the references included in this chapter [118, 144, 145]. Si et al. suggested that future work will involve harnessing the complex features found in EHRs, improving reproducibility and transparency.

Liu et al. review concentrated on representation learning of EHRs and suggested categorising these methods into statistical, knowledge-based, and graph learning methods [136]. There are four papers in Liu's review that appeared in the literature search for this chapter [145, 146, 147, 148]. Liu suggested that graphs are a practical way to represent EHRs that maintains the structural, temporal, and semantic relationships, which is not possible with other methods.

Hossain et al. review of 36 papers explored the use of EHR data for disease prediction [106]. One of their papers was included in this literature review search [147]. Hossain et al. review found that different MLs methods worked best for various clinical settings. Graph-based methods appeared to work best for Diabetes Mellitus (DM), and Hossain suggested that graph representations enable the relationships between healthcare data to be structured, enabling an understanding of connections that otherwise might be difficult to observe.

3.3 Systematic Review Methods

This systematic review follows the 2015 PRISMA protocols [149]. The completed checklist can be found in Appendix A.1. This review is registered on PROSPERO with the protocol registration number CRD42022315782.

3.3.1 Search Strategy

Synonyms for "Graphs" and "Electronic Health Records" were combined for the search strategy. Asterisk wildcards were applied to "Prognostic," "Diagnostic," and "Prediction" to expand the search. Queries targeted abstracts, titles, and keywords. Studies within this graph/ network domain were evaluated to determine if other terms cover the same concept. The search, conducted on February 27, 2023, covered MEDLINE, Scopus, and Web of Science databases. A forward citation search of review articles identified at the abstract title screening stage was conducted. Table 3.1 shows the concepts and search terms we used to run this search. The complete search strings used in each database are given in Appendix A.2.

Table 3.1: Concepts and search terms.

Concepts	Search Terms
Graphs	graph OR graphs OR graph-based OR node* AND edge* OR "knowledge graph" OR "network analysis"
Electronic health records	"electronic health record" OR "medical records systems" OR "record-linkage" OR (routine adj5 data) OR ((electronic OR link* OR compute* OR anonymi?ed) adj5 record) OR ((health OR patient OR clinic* OR medic* OR case) adj5 (record* OR data OR plan* OR chart*) adj5 (compute* OR system OR electronic OR link* OR dataset OR network)) OR EMR OR EPR OR EHR
Prognostic/ Diagnostic Prediction	predict* OR diagnos* OR prognos*

3.3.2 Inclusion Criteria

In this review, "graph" is specifically referred to as representing information in a network of nodes and edges within graph theory. The common meaning of charts/ visualisations were excluded. Included studies constructed graphs directly from individual patient-level EHR data, excluding those using aggregated population data. To assess the effectiveness of graph representation in ML-based predictions, only primary research studies that described at least one ML prediction task using EHR graph representation were considered.

Outcomes were defined as diagnostic prediction of a health condition (e.g. Heart Failure (HF) or cancer) or prognostic prediction of a health-related outcome (e.g. mortality, readmission risk, or treatment success). Studies that predicted multi-class outcomes with over ten possible labels were excluded, as statistical reporting of these models is insufficient.

Graph-based learning in healthcare has gained recent attention [1]. Considering the impact of

hardware availability on deep learning progression, papers published between 2002 and 2023 were focused to align with the release of Torch, a popular ML library framework in 2002.

Grey literature (theses, dissertations, non-peer reviewed pre-prints, and online repositories) were excluded, only full-text papers written in English or with an English translation were included.

3.3.3 Article Selection

Zoe Hancox (ZH) and Allan Pang (AP) independently conducted title/abstract and full-text screening stages, reaching a consensus on selection at each stage. Disagreements were resolved by Samuel Relton (SR), the third reviewer, and Rayyan's online software tool was employed in this process [150].

3.4 Data Extraction Methods

To perform a reproducible assessment of the included studies, two assessment frameworks were used to evaluate the RoB and identify characteristics. Both assessments were conducted independently by ZH and AP for each study.

3.4.1 Risk of Bias (RoB)

RoB was evaluated using the PROBAST tool, a framework for assessing the quality of methodologies, including RoB and applicability, in primary studies developing prediction models for diagnosis and prognosis [151].

PROBAST, developed through a consensus process with 20 signalling questions across four domains (participants, predictors, outcome, and analysis), assigns a RoB score of High, Low, or Unknown to each domain based on signalling questions. The overall RoB is determined by the worst domain score (i.e., an overall low RoB requires every domain to score low) [151]. Reviewers reached a consensus on RoB at the domain level, and a qualitative analysis was conducted for overall and domain-specific RoB across all studies. Papers with high RoB were included to highlight ongoing work in the field. The primary health outcome was focused on when multiple models are developed in a study.

3.4.2 Study Characteristics

CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) was adapted for extracting study characteristics, originally designed for primary studies on diagnostic or prognostic prediction models [152]. Qualitative analysis of the data extracted using the modified CHARMS framework revealed patterns in the identified studies. Appendix A.3 provides a detailed description of all extracted variables.

3.5 Results and Discussion

3.5.1 Article Selection

The database search yielded 1,346 papers (Web of Science (n = 410; 30.5%), Scopus (n = 633; 47.0%), and MEDLINE (n = 303; 22.5%)), with 832 unique papers. Exclusions during the title/ abstract screening were mainly due to non-predictive studies (n = 250), lack of EHR use (n = 205), and techniques not relevant to the review research questions (n = 162). Table 3.2 shows the reasons for exclusion at the title/ abstract screening stage. *Note there were papers that had multiple reasons for exclusion so the total sums up to more than the count of papers screened.* Title and abstract screening identified 37 reviews or background articles, where six papers were identified from forward citation screening for full-text screening [74, 139, 153, 154, 155, 156].

Table 3.2: Reasons for exclusion at the title/abstract screening stage. EHR=electronic health record.

Reason for Exclusion	# Excluded
Not a predictive study	250
Wrong technique	162
EHR not included	305
Wrong domain	99
Wrong type of network/graph	107
Does not explore health outcomes	104
Non-human participants	49
Pre-2002	26
Wrong publication type	26
Background article	35
Wrong study design	20
Wrong population	4
No use of graphs	8
Tutorial	1

Full-text screening identified 27 papers for data extraction. The PRISMA flowchart in Figure

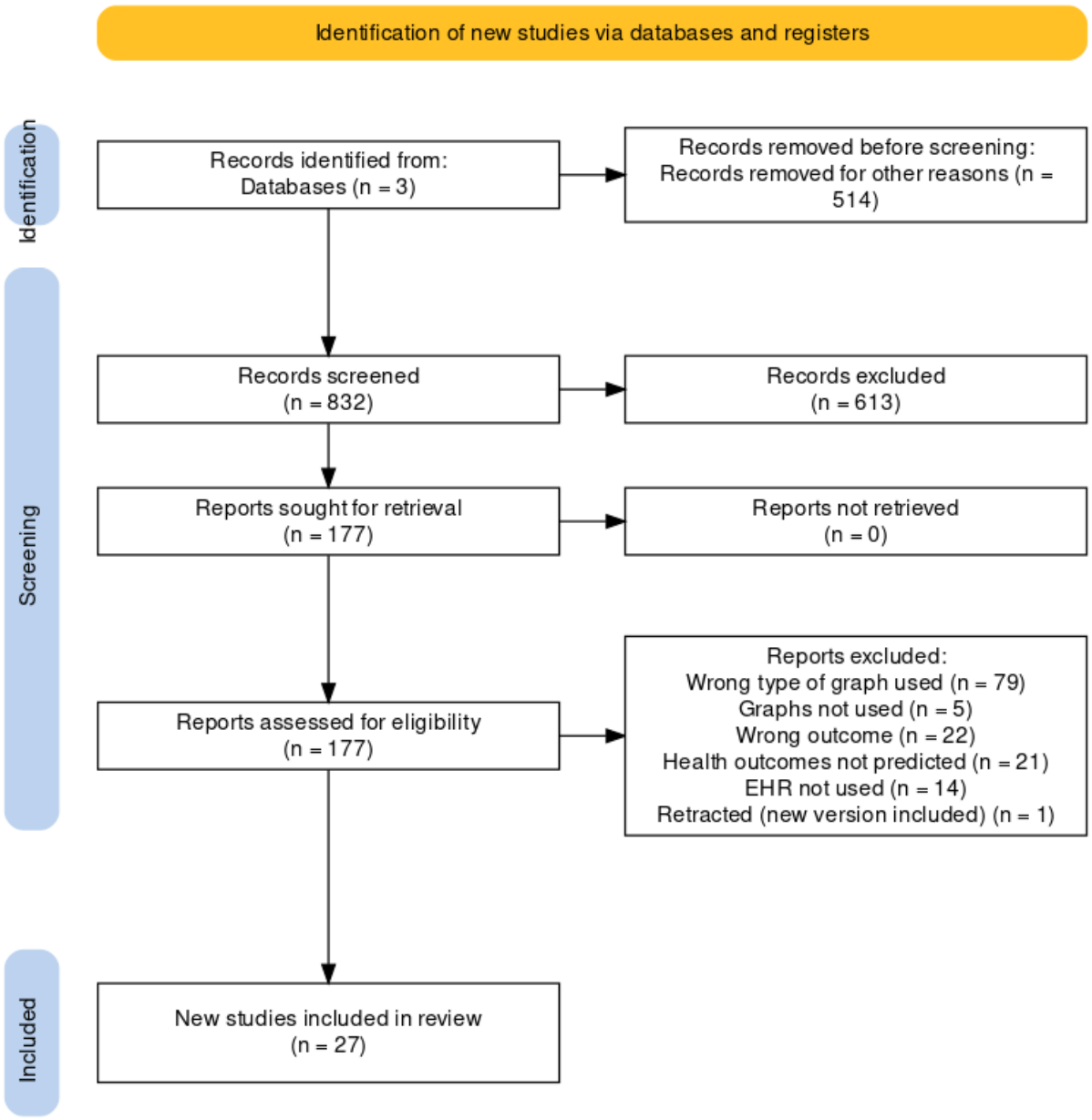


Figure 3.1: PRISMA flow diagram of search strategy. Figure produced using [157].

3.1 provides a summary of article selection. Figure A.1 in Appendix A.4 shows details of the addition screening carried out.

Within the 4 years since Schrodt et al. performed their review, the number of papers published using individual graph representation of EHR data with health predictions increased from 3 (Schrodt et al. papers) to 18 (the applicable papers from this review chapter). Figure 3.2 shows the number of papers that were included and excluded from each year between 1990 to Feb 2023 at each stage of the screening.

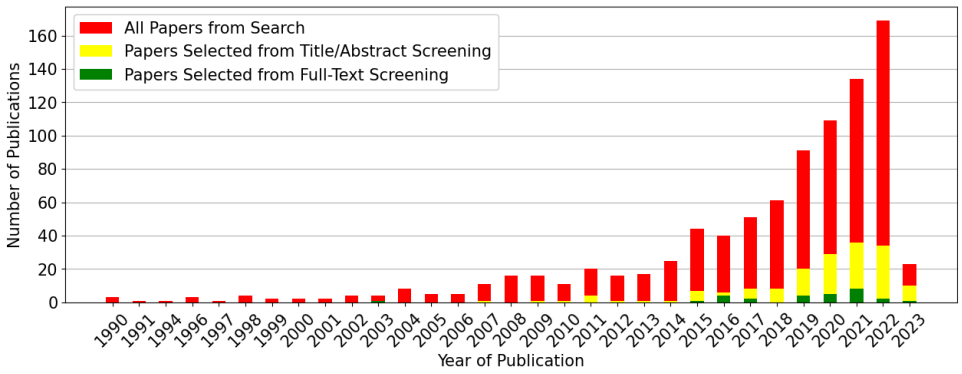


Figure 3.2: Number of papers meeting the search term criteria over the years.

3.5.2 Risk of Bias Analysis

How do biases manifest within this literature, and what impact do they have on the reliability of study results?

Table 3.3 displays the RoB and applicability assessment at the domain and overall levels, with each row representing one study. Figure 3.3 presents the breakdown of RoB levels.

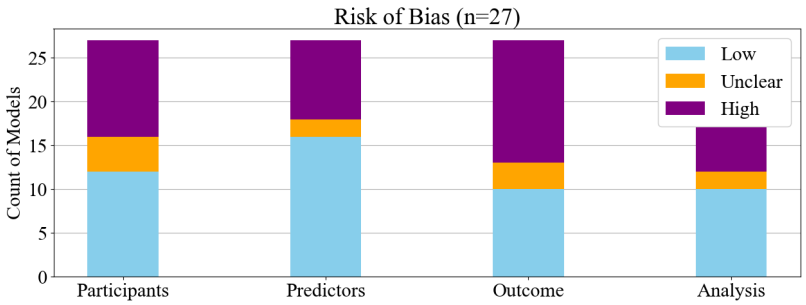


Figure 3.3: Risk of bias of the papers included for data extraction.

Table 3.3: Risk of bias and applicability table formed from following the PROBAST guidelines. H - High risk, L - Low Risk, U - Unclear risk.

	Risk of Bias (RoB)				Applicability			Overall	
Study	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	RoB	Applicability
[139]	H	H	H	H	L	H	L	H	H
[140]	H	L	L	L	L	L	L	H	L
[158]	H	L	H	H	L	U	L	H	U
[142]	L	L	L	H	L	L	L	H	L
[141]	H	L	L	L	L	L	L	H	L
[159]	L	L	L	L	L	L	L	L	L
[146]	L	L	H	H	L	L	L	H	L
[147]	H	H	U	H	L	L	L	H	L
[160]	H	H	H	L	L	L	L	H	L
[161]	H	L	H	L	L	L	L	H	L
[145]	H	L	L	H	L	U	L	H	U
[68]	L	H	H	H	L	L	L	H	L
[162]	L	L	L	L	L	H	L	L	H
[163]	H	H	H	H	L	U	L	H	U
[148]	U	H	H	H	L	L	L	H	L
[164]	L	H	H	L	L	L	L	H	L
[165]	H	H	L	H	L	L	L	H	L
[166]	L	L	L	L	L	U	L	L	U
[167]	U	H	H	H	L	L	L	H	L
[168]	H	U	U	H	L	U	L	H	U
[169]	L	L	L	U	L	L	L	U	L
[118]	U	L	H	H	L	L	L	H	L
[170]	L	L	H	L	L	L	L	H	L
[171]	L	L	L	H	L	H	L	H	H
[172]	L	L	H	L	L	U	L	H	U
[173]	L	U	H	H	L	L	L	H	L
[174]	U	L	U	H	L	L	L	U	L

Participant RoB

High RoB in the participant domain was categorised into three groups. The first group had limited information about the target population, likely introducing bias based on available data [139, 140, 141, 145, 147, 160, 161, 163, 168]. This differs from papers with unclear RoB, where no information was given about data sources or the population [148, 167], or where the control group was unclear [174]. The second group comprised papers with inclusion/ exclusion criteria leading to inclusion bias [145, 147, 158, 160, 165]. These papers selected or excluded patients based on specific characteristics relevant to the predictive context. The final group lacked control/ comparison groups to determine prediction effectiveness [118, 139].

Predictor RoB

Diagnostic codes in EHRs signify the presence of diseases and can serve as features for prediction or target labels. However, these codes do not always indicate a confirmed diagnosis. For instance, a GP may modify the rubric of a recorded code to reflect clinical uncertainty, such as "DM r/o," meaning diabetes mellitus ruled out. This highlights the importance of interpreting diagnostic codes within their clinical context. Incorrectly applying non-reproducible transformations or grouping medical events, International Classification of Diseases (ICD) codes, and medication codes poses the risk of misclassifying predictive features [68, 147, 148, 160]. Utilising tools like OpenSAFELY helps find approved code lists for appropriate patient grouping [175]. Understanding the nuances of diagnostic coding can mitigate potential biases and enhance the accuracy of predictive models.

False assumptions about the data context in EHRs can introduce inconsistencies and RoB. In acute settings, diagnoses are presumed, not confirmed, and should not be used as predictors [164]. Inappropriate imputation of missing data for ML algorithms can create unrealistic data given the clinical context and should not be used for prediction [165]. Additionally, self-reported lifestyle factors pose a potential risk of recall bias and should be interpreted with caution in predictive models [163]. However, certain lifestyle indicators, such as smoking status, are recognised as critical predictors for multiple health outcomes and are mandated to be recorded across multiple visits. While recall bias remains a concern, these factors can provide valuable insights when used in conjunction with other reliable data sources and validated coding practices.

The clinical utility of a predictive model depends on considering the timing of the prediction.

Validity requires taking into account the availability of variables at the time of the prediction. Variables only available after the event time horizon should be avoided, as they render retrospective predictions clinically irrelevant [68, 139, 167].

Two papers lacked sufficient information for RoB assessment in this domain. One lacked information about defined and assessed predictors [168], and the other lacks information about the timing of the index diagnosis [173].

These oversights indicate a lack of consideration for the clinical context in the design phase of model construction.

Outcome RoB

Approximately half of the papers showed high RoB in the outcome domain, falling into four groups: predictors shaping outcome definitions, flawed outcome assumptions, subjective outcome definitions, and poor methodology.

Consideration of the relationship between outcomes and predictors is essential. Specific predictors can unintentionally detect an outcome, such as using investigations or symptoms that are part of the diagnostic criteria (e.g., exacerbation of Chronic Obstructive Pulmonary Disease (COPD) predicting COPD onset [148]), the presence of a treatment regimen predicting subsequent diagnosis [158], or tests specific to the cancer outcome in question [68]. Acknowledging the potential masking of certain investigations when predicting disease onset should be explicit in the paper [147].

Constructing a composite outcome, like treatment failure, requires careful consideration. Some studies defined treatment failure as a patient having the same diagnostic code within two weeks [146, 164, 170]. However, this assumes patients will return to the clinician within two weeks and that the initial diagnosis is correct. Such an approach is not a formal assessment of prescription effectiveness and should not be used as an outcome.

Despite ICD coding standardisation, choosing ICD codes or diagnoses as target variables can be ambiguous, risking observer bias [161]. Some ICD codes encompass a broad range of diagnoses (e.g., N39 – Other disorders of the urinary system) or lack agreed-upon standards (e.g., E86 – Volume depletion). Subjectivity also arises in applying specific diagnostic criteria, such as Alzheimer's disease and HF [118, 172], or determining the cause of death [163].

Some papers had methodologically poor outcomes lacking predictive value in the clinical setting. Examples included papers with no prediction time horizon, rendering predictions irrelevant [118, 139, 160], or those with multi-outcome models predicting the following diagnosis or top k diagnoses [148, 160].

Analysis RoB

Around half of the papers had high RoB in the analysis domain (n = 15; 56%); this domain had the highest number of high-risk scores. Reasons for high RoB were divided into papers at risk of being over-optimistic and those with inappropriate analysis of results.

Many methodologies lacked a specified number of participants in the outcome group(s) [68, 118, 146, 147, 148, 158, 165, 171]. Without sample size calculations, it's unclear whether ML models have sufficient power to predict accurately, risking over-fitting and over-optimistic performance.

Data complexities like censoring and competing risks are not appropriately addressed or mentioned in the analysis [68, 139, 142, 145, 147, 165, 167, 168, 171, 173, 174]. Three papers lacked performance metrics or applied them inappropriately, such as using AUROC for highly unbalanced data [139, 147, 171]. One study treated alive patients differently from those who died at the end of the period [163].

Overall RoB

Only one paper, Golmaei et al. [159], had both an overall low RoB rating and low-risk applicability. Seven papers exhibited high or unclear RoB in a single domain, indicating that most literature in this area faced methodological issues across multiple domains leading to high RoB (n = 17; 63%).

The RoB findings from this systematic review align with those of related works in clinical prediction modelling. Yang et al. conducted a systematic review of clinical prediction papers from 2009 to 2019, identifying 579 predictive models [176]. Table 3.4 shows their results demonstrating that candidate predictors were the most infrequently reported (10.1%). Navarro et al. performed a systematic review of 152 clinical prediction papers (models = 522) between 2018 and 2020, focusing on trends in methodological conduct reporting [177]. Navarro found that only a minority of papers performed external validation (12.5%), hyperparameter optimisation (28.9%), or provided calibration curves (5.4%).

Table 3.4: Results from review of 579 clinical prediction models from 422 papers [176]. AUROC=area under the receiver operator curve.

Description	**Frequency (n = 579)**
Described inclusion/exclusion criteria for target population	513 (88.6%)
Reported how they handled missing data	285 (49.2%)
Target population – provided through code list	100 (17.3%)
Candidate predictors – provided through code list	59 (10.1%)
Outcome – provided through code list	103 (17.8%)
Time-at-risk – reported	489 (84.5%)
Listed candidate predictors	392 (67.7%)
Observation window for predictors provided	286 (49.4%)
AUROC reported	499 (86.2%)
Calibration plot presented	149 (25.7%)
Presented the final model completely	241 (41.6%)

This highlights a deficiency in the adherence of clinical prediction models to PROBAST guidelines. This could stem from a lack of awareness of PROBAST guidance or authors prioritising predictive model performance over clinical applicability. Addressing this gap is crucial for the effective use of these models in clinical practice.

Overall Applicability

Three papers (11%) from this search were not applicable due to using population graphs instead of patient-level EHR representations [139, 162, 171]. Nineteen papers (70%) used graphs to represent patient-level EHR data. The remaining six papers had unclear definitions of graph representations. Among them, four papers (15%) appeared to use population-based representation [163, 166, 168, 172], one paper represented graphs as single visits [145], and with one paper it was unclear about the graph representation [158].

3.5.3 Characteristics of Included Studies

Tables A.1-A.9 in Appendix A.5 provide a summary of the data extraction of the 27 papers.

Datasets and Data Sources

There were 18 different datasets included within the papers, which were broadly divided into three groups: open source, dedicated research databases, and non-public/ proprietary data. The only open source data used were the Medical Information Mart for Intensive Care (MIMIC) datasets and the eICU Collaborative Research Database (eICU) collaboration, which in this

review, were the most prevalent out of the 15 datasets (n = 10/33; 30%, n=7/33 21% respectively) [141, 145, 159, 160, 165, 166, 167, 169, 172, 174]. MIMIC is a dataset that contains patients who have attended critical care and consists of ICU short-term records, inpatient data (diagnoses and laboratory test results), and discharge summaries [178]. 14 studies used datasets from the USA, 5 from Taiwan, 4 from China, 2 from Australia, 2 from South Korea, and only 1 study employed data from the UK, with 5 additional datasets from unspecified locations. Despite the increasing application of graph-based methods in healthcare for outcome prediction, there remains a notable gap in the literature regarding the use of such techniques for predicting individual patient outcomes with graphs using UK-specific data. While existing research has explored similar approaches in other regions or on aggregated datasets, the unique characteristics of UK healthcare data, including its structure, population diversity, and healthcare delivery, remain under-explored in this context. Table 3.5 provides a summary of the datasets used.

Examples of research databases include; The Taiwanese National Health Insurance Research Database (NHIRD), Taiwan's National Death registry, the Mayo Clinic EHR, NYU Langone EHR, and the UK National Cancer Registry [68, 142, 146, 161, 163, 164, 170, 172]. These EHRs provide ICD codes with timestamps and relations to laboratory tests, treatments and clinical notes, with the addition of genetic reports in the Foundation Medicine and Mayo clinic [68].

The Geriatric Health Examination would fall into the research database category, although strictly, it is not an EHR [163]. This is due to its mandatory scheduled collection, more akin to longitudinal cohort studies. While this additional data could provide valuable insights into health prediction, its structured nature may be challenging to reproduce in routine clinical practice. However, many GP visits incorporate templates for recording specific checklists. These structured records ensure consistency in data collection across visits, demonstrating that not all primary care consultations are entirely free-form.

The majority of papers use data from non-public datasets [118, 139, 140, 147, 148, 158, 162, 167, 168, 171]. These datasets contain diagnoses with timestamped information for laboratory tests and treatments. Many of these EHRs use standardised coding systems such as ICD for diagnosis and Current Procedural Terminology (CPT) for procedures/ treatments; however, one used an undefined transformation of clinical events [148].

Three papers (11%) used simulated/ synthetic data alongside a pre-existing dataset [118, 145, 163]. Synthetic data inadequately captures complexities and relationships in EHRs, leading to

Table 3.5: Summary of datasets used in the selected papers. ICU=intensive care unit, ICD=international classification of diseases, CPT=Current Procedural Terminology, EHR=electronic health record.

Papers	Dataset	Description	Source Country
Open Sources			
[165]	MIMIC-II	Clinical data related to patient admission to ICU, diagnoses ICD-9, and lab test results. Lab tests extracted every hour from admission.	USA
[141, 159, 160, 166, 169, 172]	MIMIC-III	ICU short-term records, inpatient, discharge summaries	USA
[174]	MIMIC-IV	ICU short-term records, inpatient, discharge summaries	USA
[145, 167, 172, 174]	eICU	ICD-9 and CPT procedure codes	USA
Research Databases			
[142, 146, 164, 170]	Taiwanese National Health Insurance Research Database (NHIRD)	ICD9-CM	Taiwan
[68]	Foundation Medicine Inc and Mayo Clinic EHR	Oncology genetic reports, phenotypical data. Lab tests, diagnoses, medical and family history	USA
[161]	National registry data	-	UK
[163]	Taiwan National Death Registry	ICD-9 and ICD-10	Taiwan
[172]	New York University Langone Health	Long-term inpatient and outpatient EHRs	USA
Non-public/Proprietary Datasets			
[147, 148, 158]	Medical system from a city in North China	ICD-10 codes	China
[139, 168]	Australian healthcare system	Admission information, diagnoses, procedures (ICD-10, DRG, AN-SNAP)	Australia
[167]	Paediatric EHR data from a tertiary care hospital in China	Symptoms, medical examination information, medication codes and diagnosis codes	China
[171]	In-vitro fertilisation clinic from General Hospital in Seoul	Treatment records (age, stimulation type, use of Wallace, number of embryos transferred, symptoms)	South Korea
[140]	Private healthcare hospital admission data	ICD-10 codes and administrative data	-
[118, 158, 162, 172]	Not provided	EHR Clinical Codes	-
[173]	CardioNet	EHR Data from Seoul Asan Medical Center	South Korea

poor representation of cohorts [179]. Consequently, these results were excluded from this review as these papers appeared to demonstrate techniques rather than create predictive models for clinical settings.

Healthcare delivery can broadly be divided into inpatient and outpatient/ community care, each with distinct record structures and content reflecting the delivered care type. Outpatient records, likely sparse, offer better time coverage compared to sporadic but detailed inpatient records. When designing predictive models, these differences are crucial, as the clinical utility of predictions relies on potential levers for change in these settings. Additionally, accessing different parts of the EHR (inpatient vs outpatient) must be considered, given that typical clinical end-users lack universal access.

The distinction between primary and secondary care systems is not always evident, with many hospitals providing community services. While all papers in this study seemingly used EHRs from secondary care, details and prediction targets suggest community care records' use [68, 118, 140, 142, 146, 147, 148, 163, 164, 167, 170, 172]. This raises concern as access to healthcare records varies, necessitating clarity on data requirements for model reproducibility. Models trained on primary care data are particularly valuable because primary care is typically the first point of contact for patients in the healthcare system, allowing for earlier detection of risk factors and potential health issues. Since primary care practitioners manage a wide range of conditions and see patients over time, these models are uniquely positioned to identify patterns that could signal the onset of disease, making them crucial for preventive care and early intervention. By capturing a more complete picture of a patient's health trajectory, primary care models can support more personalised, proactive care, which is essential for reducing the burden of chronic diseases and improving long-term health outcomes.

Table 3.6 summaries the sample sizes and collection period statistics from the data used in the 27 papers. Given that EHRs have only been in widespread use for the last couple of decades, the median collection period was fairly short (5 years), reflecting the limited duration of available longitudinal data in many cases. The mean sample size, however, was notably large (148,902), suggesting that even over shorter periods, EHRs can accumulate extensive datasets. The large IQR for the sample size (ranging from 4,694 to 120,913) indicates that the scope of EHR-based studies can vary significantly. Overall, the relatively short collection periods for most studies underline the challenge of long-term data availability and accessibility in EHR-based research.

Table 3.6: Statistics of electronic health record (EHR) sample sizes and collection period from the 27 studies. IQR=interquartile range.

	EHR Year of Collection Began	**EHR Collection End**	**Collection Period (years)**	**Sample Size**
Mean	2008	2014	8	148902
Median	2010	2015	5	40517
Lower Quartile	2003	2012	3.5	4694.25
IQR	8	3.5	10	116219
Upper Quartile	2011	2015.5	13.5	120913.25
Minimum	2000.00	2001.00	1.00	132.00
Maximum	2016	2019	20	1613088

However, the ability to gather large sample sizes within short periods demonstrates the potential of EHRs to provide rich data sources for clinical prediction models.

Medical data, despite anonymisation, carries a risk of re-identifying subjects [180], countering the need for accessible datasets to verify and reproduce predictive models. The popularity of MIMIC, being freely accessible, reflects its status as a benchmark dataset for verifying predictive model performance, despite its limitations of being critical care-focused.

The recent Goldacre Review supports the scale implementation of Trusted Research Environment (TRE), offering researchers a secure environment to access medical data for model development or verification [181]. This approach provides a secure yet accessible avenue for working with anonymised EHRs.

Model Types

Sub-question a): What graph approaches are researchers taking to predict these health outcomes?

Table 3.7 shows and describes the different ML and deep learning models used with graph representations of EHRs to make healthcare outcome and diagnosis predictions. Figure 3.4 shows the model categories within the extracted studies.

RNN-based models (LSTM and GRU), excel with EHR data for handling sequential/temporal information. CNNs are employed for capturing spatial correlations. Combining RNNs and CNNs can aid in learning both temporal and spatial patterns.

Deep learning is recommended for superior performance compared to other ML methods, given

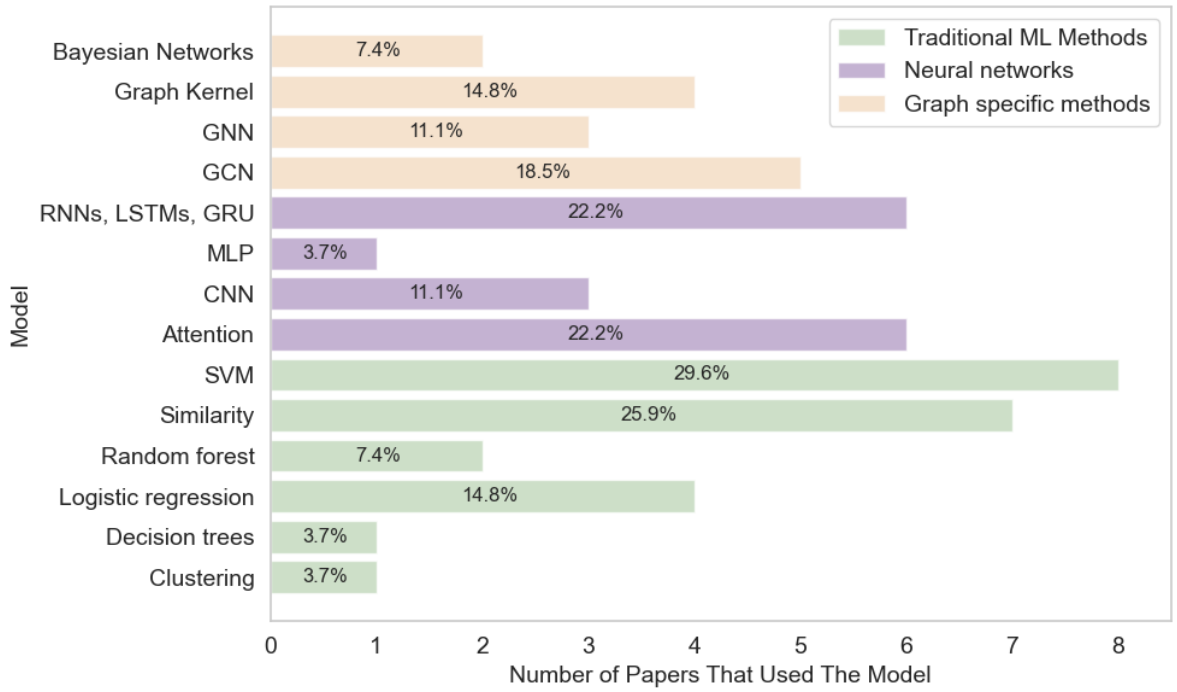


Figure 3.4: The count of models and the percentage of papers within the selected 27 papers. GNN=graph neural network, GCN=graph convolutional network, RNN=recurrent neural network, LSTM=long short-term memory, GRU=gated recurrent unit, MLP=multi-layered perceptron, CNN=convolutional neural network, SVM=support vector machine, ML=machine learning.

its capacity to capture intricate relationships. However, this complexity may not always identify uncertainties in data or model them, posing risks in healthcare settings where future predictions might suffer if the data distribution changes [10].

Figure 3.5 compares graph and non-graph models to benchmark models used for comparison.

The RoB scoring means results and performance metrics were likely biased, statistical analysis could not therefore be performed or the difference between primary and baseline models could not be assessed. Additionally, the predictive outcome would need to be identical between models for a fair comparison.

Debate surrounds the trade-off between accuracy and computational cost in ML models. Gómez-Carmona et al. demonstrated an 80% reduction in computational effort with only a 3% decrease in accuracy [182]. None of the included papers supplied information on the time taken to train their models. The exploration of resource intensity, model fitting, and prediction time for graph ML models represents a current gap in the literature. This information could be valuable in determining the feasibility of clinical implementation and optimising Pareto efficiency.

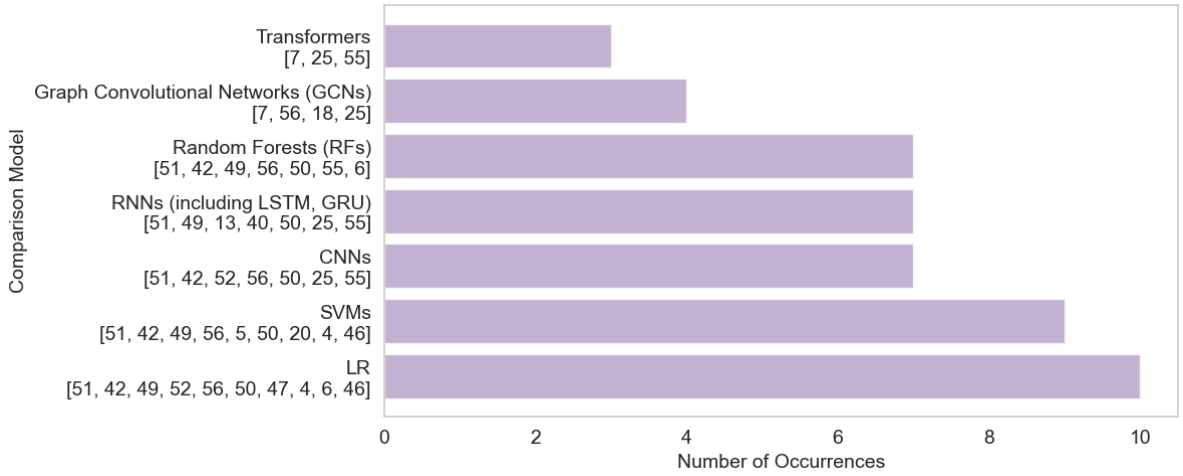


Figure 3.5: Comparison/baseline model occurrence and their associated references. RNN=recurrent neural network, LSTM=long short-term memory, GRU=gated recurrent unit, CNN=convolutional neural network, SVM=support vector machine, LR=logistic regression.

Graph Representations

Sub-question c): How are nodes and edges being utilised to perform these tasks?

Table 3.8 and Table 3.9 display node and edge types in the graphs, respectively. EHR Clinical Codes commonly served as node representations. Nodes were either homogeneous (41%) or heterogeneous (59%). Tables A.10 and A.11 in Appendix A.6 detail node and edge assignments for each model.

Graph representations allow data organisation in a non-linear path, providing explainable visualisation in a way that would otherwise be complicated to infer. Whilst minimising 'black-box' models is beneficial to provide explainability, especially in clinical settings, it is vital to consider that interpretability approaches may lead to artefacts from the learnt model rather than clinically explainable findings that should be attributed to the data [143].

Features within graphs can have various connection types to show relationships. Typical DAGs have one-to-one links (one edge can only connect to two nodes), but more complex graphs, such as hypergraphs, enable the connection of one-to-many and many-to-many links. Networks can be analysed to find relationships between node and edge components. Methods such as centrality or similarity measures can be used to look at neighbouring node contributions, connections, and structures. Data irregularity, sparsity, heterogeneity, and model opacity make modelling temporal EHR data difficult [143]; however, graphs enable us to overcome some of these challenges.

Patient data were represented using graphs to show individual patient records temporally (event data mining) or the progression between events. This was done either by including the elapsed time as the edge features or by sequentially ordering events using directed graphs without specific time intervals. Temporality is useful to include in deep learning models using EHR data, as time intervals between encounters could contain patterns not commonly known to clinicians [143]; for example, shorter time intervals may suggest poorer health.

Whilst graph representations have many advantages, they also have significant memory complexity and can take a long time to process, especially in deep learning applications.

Model Performances

Sub-question d): How do these graph approaches compare to each other? Sub-question b): How do these approaches compare to other ML, Artificial Intelligence (AI), and statistical models?

Primary ML research methodology should offer sufficient performance metrics for independent evaluation. In predictive modelling, calibration ensures that predicted probabilities accurately reflect actual outcomes, helping assess the reliability of a model's risk estimates. Discrimination measures how well a model differentiates between individuals with and without the outcome, often quantified using the C-statistic or AUROC. Decision curve analysis evaluates the clinical utility of a model by weighing the trade-offs between true and false positives at different risk thresholds to determine its practical benefit. As outlined by Steyerberg and Vergouwe, these components are essential for developing robust clinical prediction models, ensuring both statistical validity and real-world applicability in healthcare decision-making [183].

Across the papers, fourteen different metrics were provided. The most frequent metric was AUROC (70%), assessing model discrimination by comparing true positives to false positives. AUPRC followed as the second most used metric (56%), offering discriminative evaluation, particularly valuable in the presence of data imbalance. Accuracy (33%) provides a simple measure of correct predictions relative to all predictions. F1 score (26%) calculates the harmonic mean of precision and recall, preferred over accuracy in imbalanced data scenarios. Recall (26%) measures a model's ability to predict a positive outcome when present. Precision (22%) gives the positive predictive value. Specificity (7%) gauges a model's ability to predict a negative outcome when not present. Table 3.10 gives a full breakdown of all of the metrics used in these studies.

None of the papers included calibration curves or reported calibration, which helps detect overfitting by comparing predicted versus observed risk. Additionally, these papers lacked confidence intervals, a requirement for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [186].

Of the surveyed articles, some provided binary classification alone (n = 17; 63%), some multi-class (n = 2; 7%) and others gave risk/ probability scores (n = 5; 19%). A few papers had multiple predictive tasks with both binary and multi-class classification (n = 3; 11%). Hospital readmission was the most popular prediction outcome (n = 9/44; 20.5%), followed by mortality (n = 9/44; 20.5%), and then treatment success (n = 4/44; 9.1%). Table 3.11 shows all of the predicted outcomes from the 44 models and the number of times they occurred.

Table 3.12 displays models predicting mortality with AUROC or AUPRC scores. The highest AUROC score, 91.59%, was reported by Liu et al. [167]. Sun et al. [160] reported the best AUPRC score for mortality prediction at 81.34%. Figure 3.6 shows the AUROC and AUROC for the models which predict mortality. AUROC is a suitable metric to use if the dataset is balanced, however if it is not balanced AUPRC gives a better performance metric. The baseline score for AUPRC is typically determined by the prevalence of positive outcomes in the dataset. [160] had an unbalanced dataset, whilst [169] did not report dataset balance, which might explain the discrepancies between the AUROC and AUPRC scores.

Table 3.13 presents models predicting readmission with AUROC or AUPRC scores. The top-performing model for readmission prediction was reported by Golmaei et al. [159], achieving 85.8% $\pm$ 1.2 for AUROC and 84.7 $\pm$ 1.5 for AUPRC.

Table 3.14 and 3.15 display models that predicted health outcomes, excluding mortality or readmission, with AUROC or AUPRC scores. Treatment success was a frequently predicted clinical outcome.

Due to dataset and validation method variations, quantified or statistical comparisons between the papers could not be performed. However, among the three papers with low RoB predicting hospital readmission using the MIMIC-III dataset [159, 166], GNN with Bi-directional Encoder Representation from Transformers (BERT) outperformed the GCN model with attention (AUROC +3.3%, AUPRC +21.5%). The higher-performing model underwent a more rigorous 5-fold cross-fold validation validation, enhancing confidence in these results.

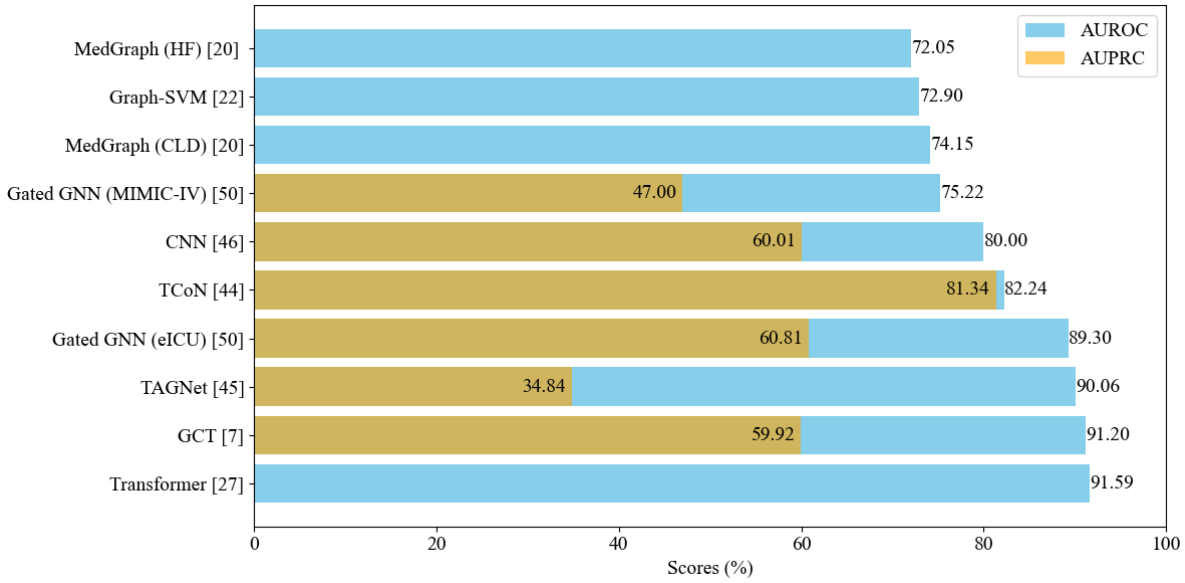


Figure 3.6: Area under the receiver operator curve (AUROC) and area under the precision recall curve (AUPRC) scores for the models predicting mortality. HF=heart failure, SVM=support vector machine, CNN=convolutional neural network, GNN=graph neural network, GCT=graph convolutional transformer.

All included papers had main models outperforming or showing equivalent results to baseline predictive performance. However, despite this being relevant to sub-question b) this improvement regularly seen might be due to publication bias favouring papers with positive improvements. The high RoB suggests likely biased results and performance metrics, preventing statistical analysis or assessment of differences between primary and baseline models. A fair comparison would require identical predictive outcomes between models. As mortality was the most common outcome to be predicted the average AUROC difference between the comparison models and the main model from each of the 8 papers and 10 models within these is shown in Figure A.2 of Appendix A.7. In most scenarios SVM and LSTM models alone had the largest difference in performance to the primary graph models.

Validation is sub-optimal with a single data split; cross-fold validation or bootstrapping is preferred for calculating standard deviations, confidence intervals, and accommodating data variations during train/ test splitting. External validation was lacking, slowing the implementation of predictive models into clinical practice. Only four papers (15%) offered links to their GitHub repositories for model availability and reproducibility [160, 162, 172, 174].

Table 3.13 shows all the models that predict readmission and provide AUROC or AUPRC scores. The model with the best performance for readmission prediction was [159], with 85.8%

± 1.2 for AUROC and 84.7 ± 1.5 for AUPRC.

Table 3.14 and 3.15 shows the performance of the models that predicted health outcomes that are neither mortality or readmission. Treatment success was the next most common clinical outcome predicted.

Table 3.7: Descriptions of the different models used within the selected papers to make healthcare predictions. RNN=recurrent neural network, LSTM=long short-term memory.

Papers	Model Type	Description
Traditional machine learning Methods		
[139]	Clustering	Grouping similar data within a dataset using 2+ variables.
[140]	Decision Trees (DTs)	A single tree that makes predictions using previous answers. This forms a series of questions in a branched shape leading to the outcome.
[68, 140, 173, 174]	Logistic regression	Linear classifier, which analyses the relationship between variables, using statistical analysis to predict binary outcomes. Note: the papers that use this method change the graph embeddings into vectorial representations.
[68, 173]	Random Forest	Multiple decision trees (DTs) trained via bagging techniques to optimise the predictive performance.
[139, 140, 158, 161, 162, 165, 168]	Similarity	Comparing 2+ samples to each other using distance or differences.
[68, 118, 142, 146, 148, 168, 170, 174]	Support vector machine (SVM)	Used for both classification and regression, SVMs find the best hyperplane to divide the data into their groups.
Neural Networks		
[145, 159, 160, 166, 167, 172]	Attention	Enabling attention to be paid to more valuable variables and reducing inefficiencies. It can also be used to show variables of importance and provide decision explainability.
[68, 141, 147]	Convolutional neural network (CNN)	Finds patterns in matrices (e.g., images, signal data) by applying filters and obtaining higher representations of the input data.
[68]	Multi-layered perceptron	Neural network which is fully connected, the connections have varying weights which enforce or weakens connections to learn the patterns from input data.
[160, 161, 162, 166, 169, 174]	RNNs, LSTM, Gated Recurrent Unit (GRU)	Takes in sequential data and keeps it in memory by taking outputs from one step to the next. This means it has connections between time.
Graph Specific Methods		
[68, 142, 145, 166, 169]	Graph convolutional network (GCNs)	Like CNNs, GCNs learn using filters over data; however, GCNs can learn directly from nodes and their neighbouring nodes.
[159, 163, 172]	GNN	Neural networks can be used on graphs to analyse nodes, edges, relationships, and layouts to make predictions.
[142, 146, 164, 170]	Graph Kernel	Convolution kernels on pairs of graphs, where the result from the convolution results in a new graph kernel.
[68, 171]	Bayesian Network	Representation of conditional dependencies between variables using directed-acyclic graphs.

Table 3.8: Node allocation types in the graphs used in the selected papers. EHR=electronic health record.

Papers	Node Allocation/Use	#
[68, 118, 139, 141, 145, 147, 148, 158, 162, 164, 167, 173]	Diagnosis	12
[142, 146, 147, 158, 160, 161, 170, 172]	EHR codes (e.g. ICD-10), but which type(s) are unclear e.g. diagnoses, demographic	8
[68, 142, 162, 163, 164, 169, 170]	Demographics (e.g. age, BMI, gender)	7
[68, 118, 148, 158, 164, 173, 174]	Medication	7
[68, 141, 145, 148, 173, 174]	Laboratory investigations	6
[145, 167, 171, 174]	Treatment	4
[68, 141, 173, 174]	Patients	4
[163, 167, 173]	Physical examinations	3
[167, 171]	Symptoms	2
[159, 162]	Clinical note representation	2
[162, 173]	Visits	2
[163]	Mental tests	1
[163]	Habits	1
[68]	Genetic data	1
[140]	Comorbidity occurrence count	1
[68]	Family history	1
[166]	Average values of word embeddings from: unique words from clinical free text or the linked Unified Medical Language System (UMLS)	1
[165]	Discretized measurements of variables at a point in time	1
[168]	EHR Features	1
[169]	Heart rate, blood pressure and oxygen saturation	1
[169]	Eye-opening and verbal response	1
[173]	Smoking	1
[173]	Echocardiography	1

Table 3.9: Edge allocation types in the graphs used in the selected papers. EHR=electronic health record.

Papers	Edge Allocation	#
[142, 146, 148, 162, 164, 170]	Time difference/elapsed between each node	6
[118, 147, 148, 163]	Temporal proximity weighting	4
[139, 140]	Number of times two diseases occurred simultaneously	2
[159, 168]	The similarity between 2 nodes	2
[146, 170]	Link to demographics (as the first node)	2
[139, 165]	Sequential directionality/ ordering	2
[139, 140]	Number of times two diseases occurred sequentially (one directly after another)	2
[169, 172]	Fully connected initially and updated by attention	2
[68, 145]	Association weighting between nodes	2
[173]	Relationship between patient and medical node e.g. edge exists between patient and smoke if the patient smokes	1
[158]	Weights higher if two medical events are more often and closer	1
[161]	Risk of disease	1
[162]	Different interactions, e.g. code to timestep	1
[166]	1) Intradocument interaction level. 2) Path lengths between entity nodes. 3) String similarities based on word overlap. 4) Cosine similarities	1
[171]	The conditional probability of a connection between 2 nodes	1
[167]	Medical relationship between nodes	1
[165]	Labelling of change of quantifiable variable (up, down or no change)	1
[160]	Linking of nodes/EHR codes happening on the same visit	1
[141]	Testing or diagnosis of a patient undertaken	1
[174]	Events happening on the same time step are linked via edge and weighting is value from laboratory test, or infusion drug. If patient took a prescription the edge weight is 1 otherwise it is 0 to the prescription node	1

Table 3.10: Performance metrics used within studies.

Metrics used	**N papers**	**References**
Area under the receiver operator curve (AUROC)	19	[68, 118, 141, 142, 145, 146, 148, 159, 161, 164, 165, 166, 167, 170, 173, 174, 184, 185]
Area under the precision recall curve (AUPRC)	15	[68, 118, 141, 145, 148, 159, 161, 162, 166, 170, 172, 173, 174, 184, 185]
Accuracy	9	[142, 146, 158, 163, 164, 168, 170]
F1	7	[148, 158, 159, 163, 165, 166, 168]
Recall	7	[148, 158, 159, 163, 165, 166, 168]
Precision	6	[158, 162, 163, 165, 168, 172]
Specificity	2	[165, 168]
Negative predictive value	1	[165]
Coverage (num predicted target diseases/ num total target diseases)	1	[158]
True positive, true negative, false positive, false negative	1	[140]
Minimum of precision and sensitivity	1	[184]

Table 3.11: Outcomes predicted within the 27 studies from 43 models. CHF=chronic health failure, CKD=chronic kidney disease, COPD=chronic obstructive pulmonary disease.

References	Outcome Predicted	N Papers
[145, 159, 162, 165, 166, 167, 168, 172, 185]	Hospital readmission (30-day after discharge, or during hospital stay (n=1))	9
[141, 145, 162, 167, 168, 169, 172, 174, 185]	Mortality	9
[142, 146, 164, 170]	Success or failure of a treatment or drug prescription	4
[158, 185]	Top k diseases	3
[118, 173, 185]	Heart failure	3
[139, 140]	Risk of diabetes	2
[174, 185]	Sepsis	2
[158]	Risk prediction of pairs of: CHF, Diabetes, CKD and COPD	1
[148]	COPD	1
[148]	Chronic heart disease	1
[68]	Type of cancer	1
[163]	Cause of death	1
[163]	High-risk patient classification	1
[167]	Acute upper respiratory tract infection	1
[118]	Prediction of heart failure related hospitalization	1
[171]	Pregnancy	1
[172]	Alzheimer's disease prediction	1
[161]	Probability of an event happening at a specific time (incl. heart failure, urinary system issues, sepsis, benign neoplasm, dehydration, heart disease, intestine disease, biliary disorders)	1
[145]	Masked diagnosis code (unspecified disease)	1

Table 3.12: Mortality prediction (binary) model performance. AUROC=area under the receiver operator curve, AUPRC=area under the precision recall curve, CNN=convolutional neural network, GRU=gated-recurrent unit, RNN=recurrent neural network, SVM=support vector machine, GCN=gated convolutional network, GNN=graph neural network, HF=heart failure, CV=Cross-fold Validation.

Paper	Models used	Dataset	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[141]	CNN	MIMIC-III	In-hospital Mortality	80.00 ± 1	60.01 ± 1	10-fold CV
[160]	Co-occurrence-aware self-attention mechanism, Time-aware GRU (T-GRU)	MIMIC-III	Mortality	82.24	81.34	Train/ val/ test 0.75:0.1:0.15 5-fold CV
[145]	Graph Convolutional Transformer (GCT)	eICU	Mortality	91.20 ± 0.48	59.92 ± 2.23	Train/val/test 8:1:1 split five times
[162]	Gaussian embedding, RNN	-	Mortality for: Heart failure Chronic liver disease	HF 72.05 Chronic liver disease 74.15	-	Train/val/test 80:15:5
[167]	Transformer	eICU	Mortality	91.59	-	Train/val/test 8:1:1
[168]	L1-SVMs, Octagonal Shrinkage and Clustering Algorithm for Regression	Australian hospital	1-year Mortality of cancer patients	72.90	-	Randomly divided into train and test sets 100 times
[169]	GRUs, attention, GCNs	MIMIC-III	Mortality in the next 24 hours	90.06	34.84	Train/val/test 70:15:15
[172]	GNN, attention	MIMIC-III	Mortality 24 hours after admission	-	71.02	Train/val/test 8:1:1
[174]	Gated GNN	MIMIC-IV	Mortality caused by HF	75.22 ± 1.52	47.00 ± 2.13	5-fold CV
[174]	Gated GNN	eICU	Mortality caused by HF	89.30 ± 0.20	60.81 ± 0.76	5-fold CV

Table 3.13: Readmission prediction (binary) models with performance metrics. CV = Cross-fold validation.

Paper	Models used	Dataset	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[159]	GNN, BERT	MIMIC-III	30-day hospital readmission	85.8 ± 1.20	84.7 ± 1.50	5-fold CV
[160]	Co-occurrence-aware self-attention mechanism, T-GRU	MIMIC-III	Readmission	74.03	72.78	Train/val/test 0.75:0.1:0.15
[145]	GCT	eICU	Readmission during the same hospital stay	75.02 ± 1.14	52.44 ± 1.42	Train/val/test 8:1:1 split five times
[165]	Non-negative Matrix Factorization	MIMIC-II	30-day ICU readmission risk	66.10	-	5-fold CV
[166]	GCN, attention, Bi-Directional Long Short-Term Memory (BI-LSTM)	MIMIC-III	30-day ICU readmission risk	82.50	63.20	Train/val/test 8:1:1
[167]	Transformer	eICU	Readmission during a hospital stay	76.14	-	Train/val/test 8:1:1
[168]	L1-SVMs, Octogonal Shrinkage and Clustering Algorithm for Regression	Australian hospital	30-day hospital readmission Acute Myocardial Infarction	63.70	-	Randomly divided into train and test sets 100 times
[118]	SVM	-	Risk of HF-related hospitalization/ readmission	73.00	67.00	Random training and testing sets
[172]	GNN, attention	eICU	Readmission prediction at discharge	-	39.86	Train/val/test 8:1:1

Table 3.14: Prediction performance of models that predict health outcomes other than mortality or readmission (1/2). CV=Cross-fold validation.

Ref	Models used	Dataset	Binary/ Multi-class	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[68]	Random Forest, Bag of Features and Node2Vec	Mayo Clinic and Foundation Medicine Inc	Multi-class	Cancer: Colon, Pancreas, Ovary, Prostate, Connective and other soft tissue, Thyroid gland, Breast, Liver, Lung	96.56	-	10-fold CV
[142]	GCN, SVM	Taiwanese NHIRD	Binary	Treatment success for: Hypertension (HTN), Hyperlipidaemia, DM	HTN 73.71, Hyperlipidaemia 74.28, DM 66.02	-	Train/ val/ test 80:10:10
[146]	K-SVM (kernel function)	NHIRD	Binary	Treatment success for: Pneumonia, Acute Otitis Media (AoM), Acute cystitis, Urinary Tract Infection (UTI)	Pneumonia 69.69, AoM 68.60, Acute cystitis 72.00, UTI 74.02	-	Train/ val/ test 80:10:10. Fine-tuned via 10-fold CV
[160]	Co-occurrence-aware self-attention mechanism, T-GRU	MIMIC-III	Binary	Disease prediction: Sepsis, HF	Sepsis 84.33, HF 76.98	Sepsis 82.33, HF 73.13	5-fold CV
[161]	Deep diffusion process, LSTM	National registry data	Multi-class	Colorectal cancer ICD-10 codes	I50 74±1.14, N39 64±0.85, A41 72±0.91, D12 69±0.53, E86 72±2.25, I25 79±0.81, K63 68±0.61, K83 69±2.17	-	-
[161]	Deep diffusion process, LSTM	National registry data	Multi-class	Stomach cancer ICD-10 codes	I50 73±0.89, N39 65±0.63, A41 69±0.65, D12 68±0.65, E86 65±1.27, I25 78±0.49, K63 65±0.77, K83 69±1.26	-	-

Table 3.15: Prediction performance of models that predict health outcomes other than mortality or readmission (2/2).

Ref	Models used	Dataset	Binary/ Multi-class	Classes/ Outcomes	AUROC (%)	AUPRC (%)	Validation Type
[148]	SVM	Not specified	Binary	Prediction of disease within 180 days: COPD, CHD	COPD 76 ± 5, CHD 71 ± 3	-	Train/test 90:10 10-fold CV
[164]	Graph kernels	NHIRD	Binary	Treatment success for: Pneumonia, AoM, Acute cystitis, UTI	Pneumonia 70.56 AoM 69.12 Acute cystitis 72.01 UTI 72.49	-	Train/ val/ test 80:10:10 10-fold CV
[167]	Transformer	eICU	Binary	Prediction of acute Upper Respiratory Tract Infection (URTI) in paediatric patients	93.34	-	Train/ val/ test 8:1:1
[118]	SVM	Not specified	Binary	HF prediction (180-day window)	72	65	Random train and test sets
[170]	Graph-kernel, SVM, attention, similarity (Euclidean distance)	NHIRD	Binary	Treatment success of: UTI, AoM, Pneumonia, Acute cystitis, HTN, Hyperlipidaemia, Diabetes	UTI 62.43±2.84, AoM 62.45±2.00, Pneumonia 60.13±2.79, Acute cystitis 61.43±1.89, HTN 73.15±1.26, Hyperlipidaemia 74.78±1.93, Diabetes 70.85±1.51	-	Train/ test split with an 80:20 ratio
[172]	GNN, attention	Inpatient and outpatient EHR data from NYU Langone Health	Binary	Alzheimer's disease prediction 12 to 24 months	-	45.80	Train/ val/ test 8:1:1
[174]	Gated GNN	MIMIC-IV	Binary	Sepsis	75.22± 1.52	47±2.13	5-fold CV
[173]	Hetereogenenous graph link prediction	CardioNet	Binary	Cardiovascular disease	72	15	Train/ val split not specified

Publication Demographics

Figure 3.7 shows the locations of the first author from each paper that used graphs for individual patient outcome prediction. The USA and China appear to be doing the most work in this area.

Out of the 27 papers, 17 were accepted in conferences (5 computer science conferences, 12 computer science + medicine conferences) and 10 were accepted by journals (6 computer science journals, 4 computer science + medicine journals). This highlights the growing intersection of computer science and medicine. This trend underscores the increasing use of computational methods, such as AI and ML, in healthcare for diagnostics, predictive modelling, and data analysis. It also reflects the recognition of medical applications in technology forums, suggesting that interdisciplinary research is advancing both fields. Overall, this integration is driving innovation in medical research and improving healthcare solutions through collaborative efforts between technologists and medical professionals. Whilst these models are primary methodological studies and have strong technical details, further encouraging groundbreaking AI technology papers into medical journals may help transition models into clinical practice by introducing models to clinicians at earlier stages to gain user feedback.

3.6 Limitations

Specific search terms were required to capture the relevant literature. Due to the limited functionality of Google Scholar, this literature database was unusable and therefore may not have fully captured potential relevant literature.

While the initial aim was to examine how graph representations of EHRs have been used, only one applicable paper with low RoB was identified, and as such this chapter turned to focus on methodological bias. Current methods leading to high RoB were highlighted with an aim to promote bias reduction in future research in this area by more careful consideration of assumptions; a must if any of these methods are to be clinically implemented.

The PROBAST tool, published in 2019, is the currently accepted guideline for assessing RoB within prediction model studies [151]. Different approaches and terminology within the field of ML mean that current PROBAST reporting may not fully critically appraise ML techniques, particularly throughout the analysis domain. There is ongoing work to address this potential pitfall, with new TRIPOD-AI and PROBAST-AI reporting guidelines anticipated, with a focus

on ML methods [187]. Given that the RoB findings in this chapter show that almost all papers were deemed high RoB within the analysis domain, it would be expected that these papers would still have high RoB under new reporting guidelines.

Data pre-processing and parameter tuning heterogeneity meant quantified comparison of graphical prediction models was not effective to recommend best approaches. Instead a qualitative summary of the papers is provided to enable researchers to make their own decisions and use these suggestions to encourage improved reporting.

3.7 Future Directions

The papers in this review focus on demonstrating the utility of graph representation in improving predictive performance rather than clinical application. This is reflected by the RoB assessment which demonstrates that the literature in this area is making assumptions that would preclude its use in the clinical environment. Many of these papers failed to consider the clinical context of their prediction. These include using predictive variables that form part of the diagnostic criteria, the poor definition of clinical outcomes or using variables that occur only in the presence of the predicted outcome. Reporting conduct will only improve if authors and the bodies accepting these papers follow TRIPOD and PRISMA reporting guidelines [149, 186].

It was expected that such false assumptions would be addressed by having input from medical experts who understand the clinical context. Only 9 (33.3%) papers had clinical input in the paper, despite the papers having clinical predictive tasks. All the papers with a clinical author had a high RoB, suggesting they did not have the expertise to understand the RoB or did not have sufficient influence during the study design process. At this intersectional space of the

Figure 3.7: Heatmap showing institution location of first author. Created using Plotly.js v2.12.1

application of computer science techniques within the domain of healthcare, better integration of medical expertise into predominately computer science teams may go some way in incorporating the clinical context and improving RoB.

None of these papers have formally explored how interpretable their analysis would be from the clinical end-user perspective and how this might change/ affect clinical decision-making. Further research into formally defining the interpretability of predictive models and their effect on actionable change would be useful for graph representation and wider adoption of ML/AI within healthcare.

The ultimate goal of developing AI solutions within healthcare is to improve clinical outcomes. As with any ML modelling technique, prediction models must demonstrate robustness in other settings through external validation but must also be understood in the clinical context. The papers included in this review focus on the predictive aspect, which allows for earlier intervention and potentially better outcomes in some contexts. A larger question needs to be answered regarding the effect of improved predictions on clinical pathways and outcomes.

From this review there are three key takeaways. 1) Researchers should consider the clinical context carefully to ensure appropriate timing, code groupings, and a reasonable relationship between the outcome and predictors for clinical utility. 2) A lot of clinical research is not currently fit for clinical use due to researchers not following TRIPOD and PROBAST guidelines. The focus needs to shift from solely enhancing predictive modelling performance to improving the clinical utility of these models. 3) Graph representations have only been used for a limited number of purposes, there is further scope to expand graph models to other tasks. Graphs infrequently depict individual-level patient representation, and when employed, predictions are confined to just six outcomes (mortality, readmission, treatment success, sepsis, Cardiovascular Disease (CVD), Alzheimer's), but graph usage could be extended to a wider range of health outcome prediction tasks such as utility or cancer recurrence.

The findings from reviewing these studies determine that methodological quality is poor, and a well-crafted health prediction paper should have the following characteristics: It should be guided by the TRIPOD guidelines, ensuring transparency and reliability, minimising bias assessed through PROBAST. It begins by defining the research question and specifying the health outcome. The methodology should outline the predictors utilised within the model and the inclusion criteria, emphasising a representative sample. Rigorous internal validation by employing

techniques like cross-fold validation and bootstrapping, is essential. The use of external validation gauges generalisability, and it is acknowledged that this may extend beyond a single paper, necessitating follow-up studies by external teams. Model development relies on robust statistical methods, accounting for predictors and their interactions, and addresses missing data and biases. Transparent reporting, including calibration curves and confidence intervals, enhances result interpretability. Recent papers discuss and demonstrate some of the methods that should be used when creating models for healthcare applications [188, 189, 190].

3.8 Conclusion

This review found 27 papers which used graph representation of EHRs for health outcome or prognosis prediction. A PROBAST analysis determined that only three papers had a low RoB. Only one paper had a low RoB that was applicable to this chapter's research question (4%). This chapter presents a narrative review of how EHR data can be represented as graphs by discussing characteristics of the methodologies, including model types, outcome prediction types, and model performances.

Models ranged from traditional ML to neural network-based models. Researchers are mainly using 4 methods (GCNs, GNNs, Graph Kernels, and Bayesian Networks) to incorporate graphs into their healthcare prediction models.

The most predicted health outcomes were mortality, hospital readmission, and treatment success. Model performances ranged across outcomes, mortality prediction (AUROC: 72.1 - 91.6; AUPRC: 34.8 - 81.3) and readmission prediction (AUROC: 63.7 - 85.8; AUPRC 39.86 - 84.7). These graph approaches outperform baseline models that use non-graph ML, AI, and statistical methods. However this may potentially be due to publication bias. Diagnosis and EHR codes are most frequently being used for graph nodes, whilst edges are being used to represent time and simultaneous disease occurrence. Out of the 3 low RoB models, the GNN with BERT model had the best performance for hospital readmission prediction [159]. In the high RoB papers, the TCoN model (GNN with attention) had the best AUPRC performance.

Graph-based representations using EHRs, for individual health outcomes and diagnoses is an area ripe for exploration but require further knowledge building before results are applied clinically. Graph representations appear to be useful in dealing with the sparsity of EHRs, by

retaining structure and temporality. The simplicity of these graphs is also well suited for ML models for predicting health outcomes. Whilst mindful of publication bias, the technique of graph representation appears to improve predictive performance compared to baseline ML methods in multiple fields of medicine, suggesting the potential for universal application.

The high RoB suggests that authors do not use or are unaware of TRIPOD and PRISMA reporting guidelines. This may change with the publication of TRIPOD-AI and PROBAST-AI [187], which are specific to AI/ ML methods and may become a requirement for publication to conferences/ journals in this field. These efforts are not insurmountable with a proper study design that incorporates clinical context, which will lead to suitable models within the clinical setting.

In conclusion, the systematic literature review highlights several significant gaps in the current research landscape, which this thesis aims to address. First, few studies employ UK-based datasets, with the NHS presenting unique structural and procedural differences from healthcare systems in other countries, suggesting that findings from international datasets may lack applicability within the UK context. Second, limited research utilises graph-based methods or patient-specific graph representations derived from EHRs to predict health outcomes, despite the potential of these methods for capturing complex relationships in patient data. Third, primary care data, a crucial component given its role as the first point of contact in patient care, is absent in the datasets used in the reviewed studies, raising concerns about the comprehensiveness of existing models. Additionally, no studies provide insights into resource demands, such as the time required for model training and inference, leaving questions about the scalability of these approaches in real-world clinical settings. Lastly, none of the reviewed papers incorporate a clinical or end-user perspective on model interpretability, which is essential for ensuring that predictive models are useful, understandable, and trusted by healthcare providers. Addressing these gaps, the following chapters of this thesis will explore these under-researched areas, aiming to advance the development and application of predictive healthcare models that are both practically implementable and clinically relevant.

Chapter 4

TG-CNN Methodology

4.1 Introduction

As previous chapters have demonstrated, temporal medical data is currently underutilised for patient prediction despite patient record data continuing to accrue rapidly and pressures on health services requiring relief. In this chapter, the design and methodology of the TG-CNN model, which is able to take events with irregular time intervals, is introduced and discussed.

Due to not yet having access to a medical dataset at the time of writing this chapter, an open access MOOC dropout dataset was used. Whilst this dataset was not health-related it served as a useful dataset to initially construct the model around and enabled the exploration of student dropout prediction using MOOC temporal clickstream data.

MOOCs allow people to study and learn a wide range of material wherever and whenever they choose [191]. The COVID-19 pandemic caused a significant rise in the number of online courses available; yet MOOC retention is lower and student dropout is higher than in-person courses [185, 192]. The MOOC dataset used comprises of a temporal sequence of student clickstream actions and subsequent dropout information. Predicting dropout based on clickstream data could enable identification of behaviour patterns prior to dropout, to target interventions designed to encourage course completion [193, 194].

Graphs are useful for capturing object interactions, for example where nodes may represent people and edges depict messages from one person to another. Convolutions applied over graph structures have been shown to learn effectively in various tasks [195, 196]. The model presented

in this chapter uses temporal graphs, a three-dimensional CNN, and LSTM units to analyse clickstream data (user actions) by integrating the elapsed time between events to enhance the predictive accuracy for student dropout. Convolutions within CNNs slide a filter over input data, such as an image, to extract features like edges, textures, or patterns. The filter multiplies its weights with the input values it overlaps, producing a feature map that highlights specific characteristics. This process helps the network detect spatial hierarchies and patterns while reducing the input's size and complexity. The methodology automatically extracts informative features from patterns within the ACT MOOC data, utilising time between clicks to gain insights. The model's effectiveness is assessed against baseline models and recent methods in the literature.

In later chapters of this thesis the application to medical data and the addition of explainability of these models for clinical interpretation is discussed. With the long-term aim of producing a clinical decision tool that can promote trust in the model, aid discussion between patient and clinician, provide personalised medicine and protect patient safety.

4.2 Methodology Outline

The model described in this chapter is able to handle data that is irregularly sampled in time, which RNNss and LSTMs alone are unable to achieve. This model automatically generates informative data features, evaluating events within their context to alleviate reporting bias, whilst incorporating the elapsed time between events to improve predictive power.

4.2.1 Dataset

The MOOC dataset[1] is composed of a temporal sequence of timestamped student actions and subsequent dropout information. The MOOC dataset consists of 7,047 users, with a dropout rate of 57.7%. There are 97 potential clickstream actions a student can take.

Time was given in seconds from the first interaction a user makes with the online course. In total over 411,749 interactions were captured, with the most actions taken by one user totalling 505.

The most commonly reported clickstream action was action 8, occurring 19,474 times, whilst the action 93 was the least common action appearing 87 times.

[1]Stanford Network Analysis Project - `https://snap.stanford.edu/data/act-mooc.html`

The average elapsed time between clicks was 816,744 seconds (9.46 days), with a minimum of 0 seconds, a median of 644,820 (7.46 days) seconds and a maximum of 2,517,315 seconds (29.14 days). 0 seconds was the most common elapsed time between clicks.

4.2.2 Data Preparation and Modelling Approach

The approach taken for this prediction task was to turn a sequence of clickstream events into a temporal multigraph and formulate dropout prediction as a binary graph classification task, where each student has an individual temporal graph to be classified.

In particular, the $n = 97$ possible actions formed the nodes of this graph and the temporal edges captured the elapsed time between actions. This can be stored in a 3D tensor $G(i, j, k) = t_k$ where $i, j \in \{1, \ldots, n\}$ are nodes in the graph, and t_k is the elapsed time for the kth edge in the temporal graph. The elapsed time is the time between the previous click to the next click action, given in seconds.

If a student had less than two clicks their data was excluded from modelling, as this method requires at least two events at two separate times to form an edge of a graph.

In clickstream data, no single action is crucial for effective modelling, so missing data is typically treated as irrelevant rather than as a problem to address. The data is inherently redundant, as users often perform similar actions. This redundancy makes it impossible to determine exactly where a missing click occurred in a sequence, making accurate inference of missing data impractical. Attempting to impute missing actions could introduce inaccuracies, as it would require assumptions about the sequence that may not align with real user behaviour. Attempting imputation, either through random insertion or pattern-based inference, may degrade model performance by introducing noise. The TG-CNN model, effectively handles sequence patterns as they are, implicitly adjusting to the data's natural gaps without requiring explicit imputation. This approach avoids the potential pitfalls of adding artificial data points, allowing the model to capture authentic user behaviour patterns more reliably.

For this particular task the most recent 100 actions of each user were used to reduce computational burden. Tensors were front-padded where sequences contained less than 100 clicks to ensure the most recent actions were always at the end of the 3-tensor. With 97 potential click actions this means that each sample 3-tensor is size $97 \times 97 \times 100$. An example of how a sequence of 5 events with 4 possible actions would be converted to a 3-tensor is shown in Figure 4.1.

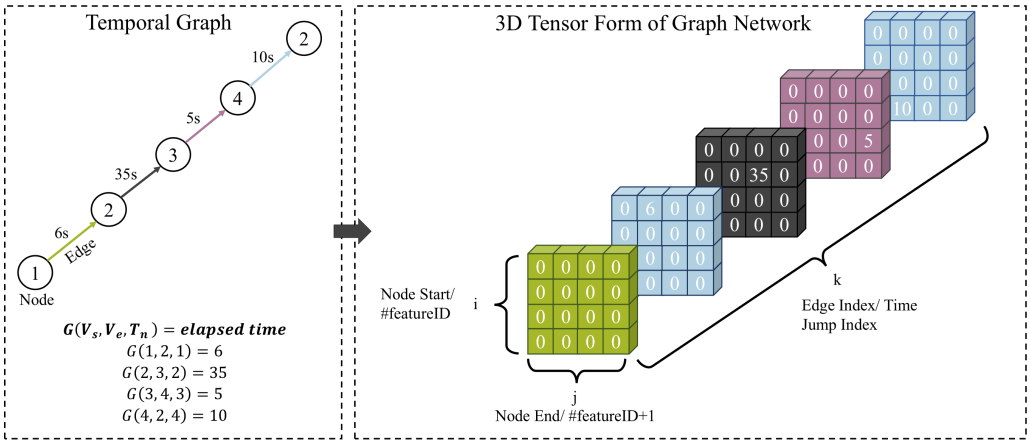


Figure 4.1: Graph network visualisation showing connections between actions completed by a user in both graph and tensor form. This example has only 4 possible actions, so is much smaller than the $97 \times 97 \times 100$ tensor that was used in this project.

It is assumed that events that happen with less time elapsed between them are more likely to be related to each other, for this reason a modified version of the elapsed time is stored which measures the proximity of two events occurring. The graph is stored with the transformation $G(i, j, k) = \exp(-\gamma t_k)$ where γ is a trainable model parameter and t_k is the elapsed time. If two events happened simultaneously then $G(i, j, k) = 1$, whereas if there was a large gap in time between two events then $G(i, j, k)$ would score closer to 0 (Figure 4.2). Unmodified elements which show no interaction between vertex pairs at a set time were set to 0 in the 3-tensor.

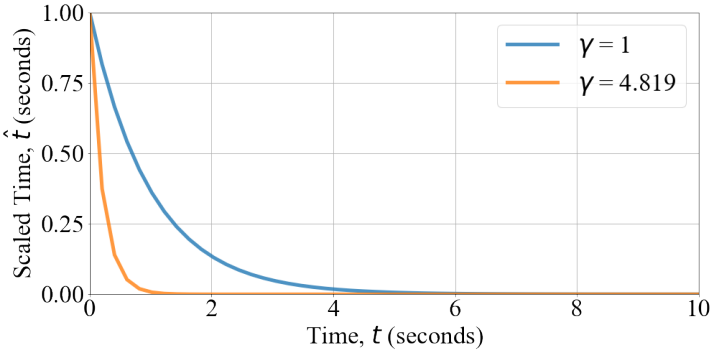


Figure 4.2: Scaling of the time ($\exp(-\gamma t)$) when $\gamma = 1$ versus when $\gamma = 4.819$.

Incorporating γ in the model has two main benefits: 1) Actions taken in quick succession or simultaneously have a value close to 1 and actions with a greater temporal gap are closer to 0. Events that are not directly adjacent in sequence are set to 0. This allows the temporal graph to be stored as a sparse 3-tensor, saving significant memory in the representation of the data. 2) Elapsed time can be rescaled to avoid extreme values in the neural network and potential

under/overflow when using half-precision arithmetic.

In particular, suppose that $\gamma = 0.01$ and event 12 is connected to event 34 via a temporal edge at time point 21 with an elapsed time of 14 days. The value at $(12, 34, 21)$ of the 3-tensor (referred to as $\mathbf{G}$) would be set to $\exp(-0.01 \times 14) \approx 0.87$. See Figure 4.3 for an example graph.

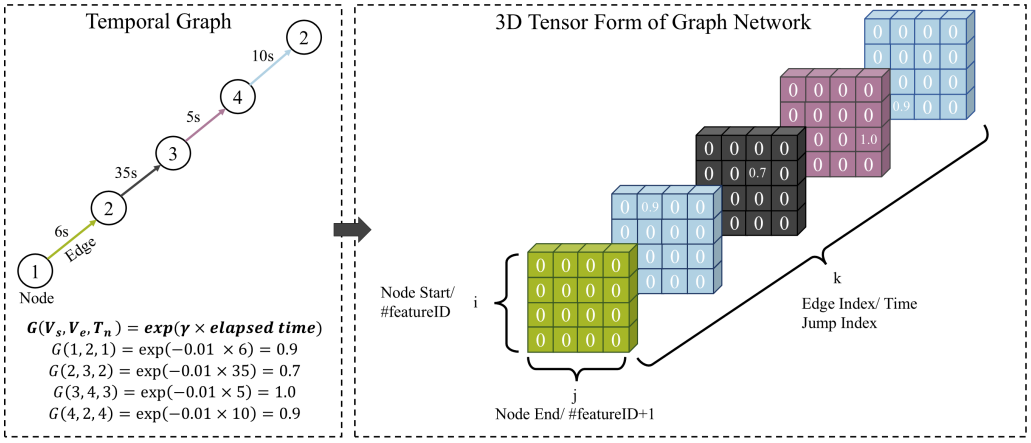


Figure 4.3: Graph network visualisation showing connections between actions completed by a user in both graph and tensor form with exponential scaling and $\gamma = 0.01$ function included.

A neural network architecture was created which applies convolutions over the time axis of the tensor, to discover subsets of the data which are similar to the patterns learnt.

In this project's preliminary work, a standard layer from the PyTorch package (Conv3d) was used to perform 3D convolutions across the temporal (i) axis of the dense 3-tensor using 3D filters (see Code 4.1 for details on the code). This is similar to standard convolutions used in CNNs for image processing.

Code Listing 4.1: TG-CNN Model created using PyTorch package.

```
class Dense3DConv(nn.Module):
    def __init__(self):
        super(Dense3DConv, self).__init__()
        self.gamma = nn.Parameter(torch.rand(1, device=device))
        self.conv3d = nn.Conv3d(in_chans, out_chans, kernel_size=(
                                                num_time_steps,97,97), stride=1
                                                , bias=False)
        self.batchnorm = nn.BatchNorm3d(out_chans)
        self.lstm = nn.LSTM(input_size = out_chans, hidden_size= lstm_h,
                                                batch_first = True)
        self.flat = nn.Flatten()
```

```
        self.fc1 = nn.Linear(conv_D_out * lstm_h, linear_size)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(linear_size, 1)
        self.sig = nn.Sigmoid()
        self.dropout = nn.Dropout(p=drop_val)

    def forward(self, x):
        x = x.to(device)
        gamma = ((gamma_max-gamma_min)*self.sig(self.gamma) + gamma_min)
        scaled_matrix = torch.where(x==0, x, torch.exp(-gamma*x))
        out = self.conv3d(scaled_matrix)
        out = self.batchnorm(out)
        out = torch.squeeze(out, 4)
        out = torch.squeeze(out, 3)
        out = self.relu(out)
        out, cell_state = self.lstm(torch.transpose(out, 1, 2))
        out = self.flat(out)
        out = self.dropout(out)
        out = self.fc1(out)
        out = self.relu(out)
        out = self.dropout(out)
        out = self.fc2(out)
        return out
```

As an example, consider that a maximum of $n_t = 100$ time steps were used with 1,000 potential events, making the 3-tensor input size equal to $1000 \times 1000 \times 100$. A filter F (with a learnt pattern) with shape of $1000 \times 1000 \times 3$, means that it convolves over 3 time points simultaneously with one degree of freedom (only moves in one axis). This filter would slide over the time points as shown in Algorithm 1. The output provided after these convolutions depends on the similarity between the section (subgraph) of events and the pattern represented by the filter F.

Algorithm 1: Temporal convolution F being applied to the sparse 3-tensor G representing the temporal graph of a patient EHR.

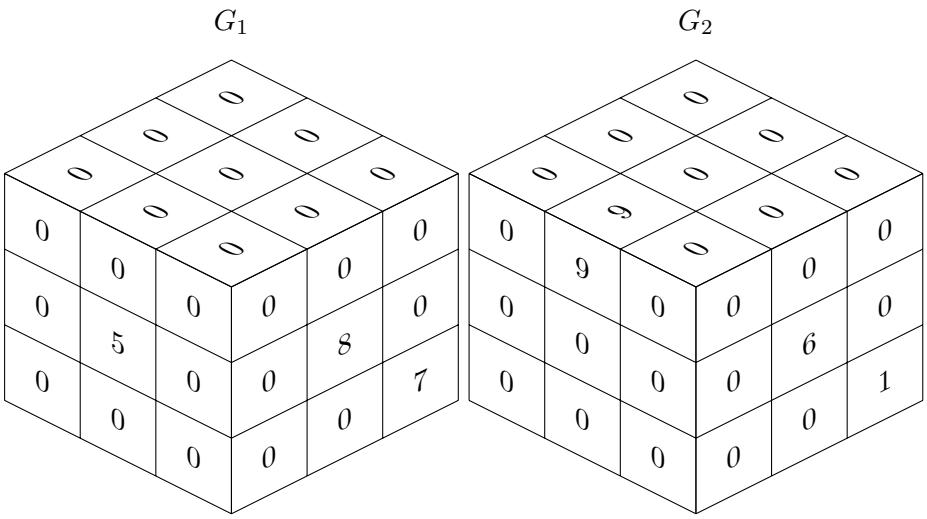
Data: G of size $1000 \times 1000 \times 100$ and F of size $1000 \times 1000 \times 3$.
Result: Vector v of length 98.
for $t = 1, \ldots, 98$ **do**
  $v(t) = \sum_{s=0,\ldots,2} \sum_{i,j} G(i, j, t + s) \times F(i, j, 1 + s)$;
end

In converting graphs to tensors and then converting these to sparse tensors, typical CNNs from PyTorch do not work. As the standard `Conv3d` layer from the PyTorch package could only perform 3D convolutions on dense 3D matrices, the graphs were loaded into the model as sparse tensors and then each batch was converted into dense tensors before feeding into the `Conv3d` layer to reduce memory usage. This reduces the RAM required; however, this could be improved upon further to improve training speed. This approach helps to reduce memory usage by only making the data dense when needed. However, optimising this workflow further could potentially improve training speed by reducing the frequency of dense conversions or by exploring alternative sparse-compatible operations.

To speed up modelling training and allow larger sample sizes to be used, the model was converted using the Tensorflow package instead. Tensorflow (version 2.8.0) was used to create a custom CNN Keras layer which utilises sparse linear algebra. The temporal graph representations can be efficiently stored as sparse 3-tensors, where the elapsed time are the values and the clickstream actions are the indices. During model training these sparse 3-tensors are batched together forming 4-tensors.

Here the steps used to perform the sparse 3D convolutional function over a 4-tensor are described with dummy examples:

In this example, graph 1 (G_1) ($3 \times 3 \times 3$) and graph 2 (G_2) ($3 \times 3 \times 3$) are given as tensors:



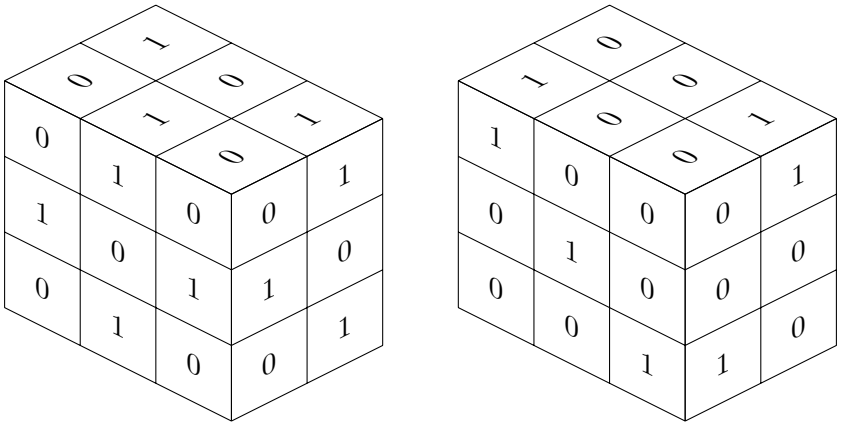
Which can also be represented as sparse tensors:

$\mathbf{G}_1$	$\mathbf{G}_2$
Values: $[5, 8, 7]$	Values: $[9, 6, 1]$
Indices: $[[1, 1, 0], [1, 2, 1], [2, 2, 2]]$	Indices: $[[0, 1, 0], [1, 2, 1], [2, 2, 2]]$

Together these two matrices can be 'batched' together by adding an extra dimension on axis 0 to produce a 4-tensor of shape $[batch\_size, num\_nodes, num\_nodes, timesteps]$ which in this case would be $[2, 3, 3, 3]$.

Two filters are used in the example depicted below (filter 1 (F_1) and filter 2 (F_2)).

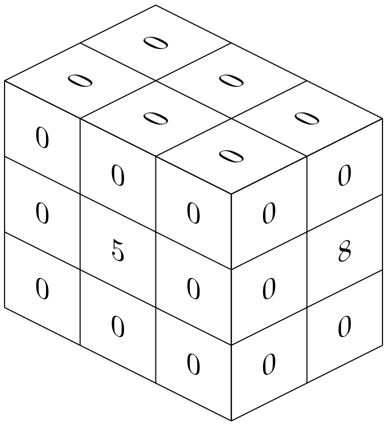
F_1 F_2



F_1 and F_2 are filters of size 2 as they cover two timesteps. The *num_nodes* for the filters will always be the same as the *num_nodes* of the input graphs as slices are only taken in the temporal direction.

First, slices of the graphs are taken, slicing to the same size as the filters. For example, slice 1 of graph 1 (referred to as G_{1s1} from now on) would be as follows with shape $[batch\_size, num\_nodes, num\_nodes, filter\_size] = [2, 3, 3, 2]$:

G_{1s1}



Note that in this example, the filter would only slide over the graph twice as the number of slices corresponds to $timesteps - filter\_size + stride = 3 - 2 + 1 = 2$.

Next, the graph slices are each flattened into a row vector. Multiple flattened graphs can be batched together by stacking them to create 2D matrices. Let $G_{i,j}$ be slice j of graph i, then the flattened graph matrix is $G = [G_{1,1}, G_{1,2}, G_{2,1}, G_{2,2}]$ and the flattened filter matrix is $F = [F_1, F_2]$.

Flattening the slices enable matrix multiplication to be performed more efficiently. The function `tf.sparse.spare_dense_matmul(graph, filter)` requires the graph to be sparse and the filter to be dense. Without `tf.sparse.spare_dense_matmul(graph, filter)` element-wise multiplication and then summation would need to be carried out on the tensors.

Continuing with the example, the flattened dense graph slices and filters would be vectors:

$G_{1s1} = \langle 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 5\ 0\ 0\ 8\ 0\ 0\ 0\ 0\ 0\ 0 \rangle$

$G_{1s2} = \langle 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 8\ 0\ 0\ 0\ 0\ 0\ 0\ 7 \rangle$

$G_{2s1} = \langle 0\ 0\ 9\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 6\ 0\ 0\ 0\ 0\ 0\ 0 \rangle$

$G_{2s2} = \langle 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 6\ 0\ 0\ 0\ 0\ 0\ 0\ 1 \rangle$

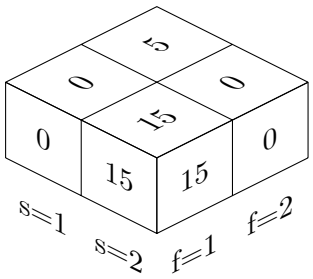
$F_1 = \langle 0\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1 \rangle^T$

$F_2 = \langle 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0 \rangle^T$

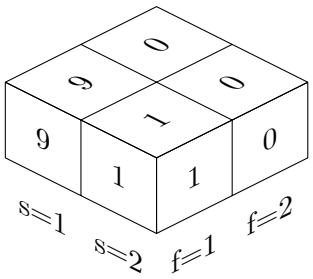
Performing `tf.sparse.sparse_dense_matmul(graph, filter)` iteratively with these graph slices and filters (as shown in Figure 4.4) would result in the following output tensor with

shape $[batch\_size, num\_filters,\ 1, timesteps - filter\_size + 1] = [2, 2, 1, 2]$:

Graph 1 Output



Graph 2 Output



The Tensorflow code for the custom 3D CNN layer, with one degree of freedom, can be seen in Code Listing 4.2 alongside the summary Algorithm 2.

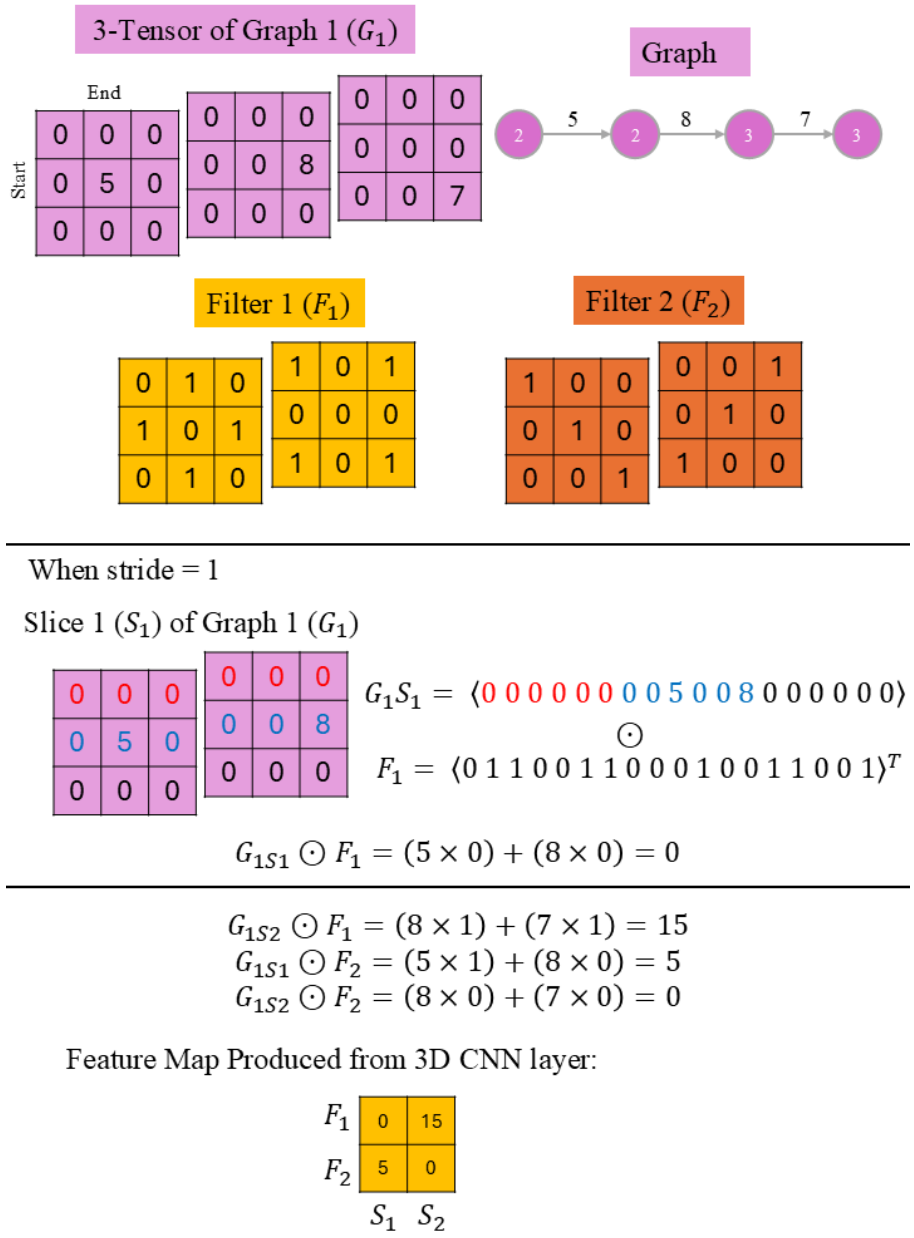


Figure 4.4: Example of how the output from the 3D CNN layer is calculated using element wise multiplication and summation.

Algorithm 2: 3D convolution applied to sparse graphs across multiple time steps.

Data: $input\_graphs$ of size $[batch\_size, num\_nodes, num\_nodes, time\_steps]$, filter w, and $filter\_size$.
Result: Tensor g of size $[batch\_size, num\_filters, 1, 1, output\_time\_steps]$.
Initialize g with the first slice of $input\_graphs$;
Reshape initial slice to a column vector and perform sparse-dense multiplication with w;
Expand dimensions of g to make it 4D;
for $k = 1, \ldots, time\_steps - filter\_size + 1$ **do**
    Extract slice of $input\_graphs$ from k to $k + filter\_size$;
    Reshape slice and perform sparse-dense multiplication with w;
    Expand dimensions and concatenate result to g;
end
return g;

Code Listing 4.2: Extract of Tensorflow (Python) code to perform 3D convolutions over multiple graphs in parallel.

```
k = tf.constant(1, dtype=tf.int64)


# Get initial slice from all graphs
# 'input_graphs' is a 4D tensor of sparse graphs
# input shape = [batch_size, num_nodes, num_nodes, timesteps]
# Reshaping so the tensor becomes a column vector
g = tf.sparse.reshape(
        tf.sparse.slice(input_graphs,
                        [0, 0, 0, 0], # start
                        [input_graphs.dense_shape[0], # size
                         self.num_nodes, self.num_nodes, self.filter_size]),
                         [-1,1]) # to column vector


# Reshaping to have shape = [batch_size, num_nodes*num_nodes*filter_size]
g = tf.sparse.reshape(g, [input_graphs.dense_shape[0], self.num_nodes*self.
                                            num_nodes*self.filter_size])


# Matrix multiplication of the first slice of the graph (sparse) and
# the weights (dense) giving shape [batch_size, num_filters]
g = tf.sparse.sparse_dense_matmul(g, self.w)


# Change g from 2D tensor to 4D tensor by adding 2 empty dimensions to the end
g = tf.expand_dims(tf.expand_dims(g, 2), 2)


# Loop slides a 'window' of the filter_size in the temporal axis to get slices
```

```
# Concatenate builds the output onto the initial slice in the '3rd' dim
def in_loop(k, g):
    g = tf.concat([g, tf.expand_dims(
            tf.expand_dims(
                tf.sparse.sparse_dense_matmul(
                    tf.sparse.reshape(
                        tf.sparse.reshape(
                            tf.sparse.slice(input_graphs, [0, 0, 0, k],
                        [input_graphs.dense_shape[0],
                         self.num_nodes, self.num_nodes, self.filter_size]),
                         [-1,1]),
                        [input_graphs.dense_shape[0],
                        self.num_nodes*self.num_nodes*self.filter_size]),
                        self.w),2),2)], 3)
    return k + 1, g # k = counter

# while_loop allows in_loop to be performed in parallel
_, g = tf.while_loop(lambda k, g: k < self.time_steps-self.filter_size+1,
                     in_loop, [k, g], shape_invariants=[k.get_shape(),
                     tf.TensorShape([None, None, None, None])],
                     parallel_iterations = self.parallel_iter)
return g
```

A filter should ideally approximate a subgraph, so that a students clickstream is close to a given filter if it is expected that their actions will follow a common path. This is enforced with regularisation. ℓ_1 (LASSO) regularisation is a shrinkage method using Manhattan distance to reduce all the weights, this enables feature selection by sending the less important weights closer or to zero. During network training ℓ_1 regularisation is applied to the filter to prevent overfitting and encourage sparsity. As ℓ_1 regularisation is calculated by taking the sum of absolute value of weights, this method is robust to outliers. This regularisation was applied by summing the loss and the ℓ_1 norm: $loss = loss + \lambda \sum_{i=0}^{n} |w_i|$ where n is the number of filters in the dataset, w_i represents each element for the ith filter and λ is the ℓ_1 regularisation strength.

4.2.3 Model Architecture

LSTMs assume a constant elapsed time between sequence elements, an issue which has received some attention in the literature [192, 197]. The TG-CNN approach offers an alternative for-

mulation of this problem (including variable time dilation), which can model more complex temporal links.

The initial TG-CNN model is shown in Figure 4.5. The 3-tensor input of size $97 \times 97 \times 100$ is fed into the 3D CNN layer, which extracts information on the sequence of actions and scaled elapsed times between actions. The outcome from the 3D CNN layer is flattened and is then passed through a batch normalisation function which speeds up convergence, followed by a ReLU activation function, before proceeding through the LSTM. The output of the LSTM has dropout applied, passing the hidden features into a Fully Connected Layer (FCL). Dropout and a ReLU are then used again before a final FCL and final output layer. A sigmoid function is then used to map the outcome between 0 and 1 for binary classification: $\sigma(x) = \frac{1}{1+e^{-x}}$. Binary cross entropy logits loss is used to measure the difference between the predicted probability and the true probability. Adam optimisation with L2 regularization was used to smooth oscillations during training. This implementation also utilises a learning rate scheduler, multiplying the learning rate by 0.9 with an exponential decay each 10,000 steps. Early stopping was used with a patience of 50, which checkpoints the model when the validation loss decreases, interrupting execution when the model gets stuck in a local minima.

The 3D CNN component has the ability to capture short-term temporal patterns of user actions, whilst the LSTM can cover longer-term associations.

RNNs may be used to improve predictive performance, as the convolution function provides a sequence of similarity scores, the RNN layer can find temporal features of these scores in sequence.

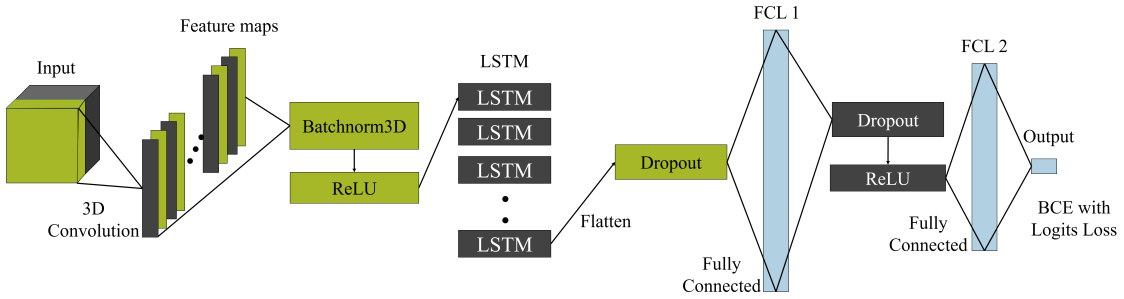


Figure 4.5: TGCNN model architecture.

A secondary ("multi-stream") architecture was experimented with to search over different granularities in time, for example one stream may look at patterns evolving over weeks whilst another stream could look at months or years. This comprised of a second branch consisting of a 3D

CNN and LSTM using filters with a stride of 2, whereas the original 1-stream had a stride of 1. The output of the two independent streams were concatenated after two FCLs, which were then followed by two further FCLs.

To assess the effectiveness of model components, an ablation study was performed to experiment with the following architectures:

- **TG-CNN w/o elapsed time.** TG-CNN is trained with elapsed time removed, leaving only the sequence of events, with all connected nodes receiving a 1 instead of elapsed time.
- **TG-CNN w/o exp.** TG-CNN is trained without the additional of an exponential function applied to the input data to scale time.
- **TG-CNN-multi-stream.** TG-CNN is trained with an additional branch to integrate a coarse and fine stream of convolutions over the graphs in parallel.
- **TG-CNN-w/o γ.** TG-CNN is trained without a trainable γ value, instead γ is set to 1.
- **TG-CNN-w/o LSTM.** TG-CNN without an LSTM layer.
- **TG-CNN-with elastic net.** TG-CNN with both ℓ_1 and ℓ_2 regularisation.

To train these models the N8 Bede machine based at Durham University: an IBM Power 9 system with NVIDIA V100 GPUs, was used initially. And then Torch version 1.7.0, Tensorflow 2.8.0, NumPy 1.19.2, Pandas 1.2.4, Scikit-Learn 0.23.1, and CUDA 10.2.89 were used on a desktop with a NVIDIA RTX 3090.

4.2.4 Model Evaluation

The following metrics were used in this thesis to evaluate the TG-CNN models:

1. *True Positive (TP)*
2. *True Negative (TN)*
3. *False Positive (FP)*
4. *False Negative (FN)*
5. *Confusion matrix:* A grid which shows the number of all positive and negative outcomes.

		Actual Values: Positive	Actual Values: Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

6. *Classification accuracy:* The number of correct predictions given correctly in the sample.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

7. *False Positive Rate (FPR):*

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

8. *Recall (Sensitivity or True Positive Rate (TPR)):* How well the model predicts a positive outcome when that positive outcome is truly present.

$$Sensitivity/Recall = \frac{TP}{TP + FN}$$

9. *Specificity:* How well the model predicts a negative outcome when that negative outcome is not present.

$$Specificity = \frac{TN}{TN + FP}$$

10. *Precision (positive predictive value):* The proportion of correctly predicted positive outcomes out of all outcomes predicted as positive.

$$Precision = \frac{TP}{TP + FP}$$

11. *F1 Score:* The harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

12. *AUROC:* The area under the curve of the TPR vs the FPR plot, showing how the model performs. It is computed by integrating the TPR as a function of the FPR across all

threshold values:

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

This integral represents the cumulative AUROC, typically calculated using the trapezoidal rule. The AUROC usually gives a value between 0.5 and 1, where 0.5 means the model has no discriminating ability (as good as random chance) and 1 indicates it has the ability to perfectly discriminate between different outcome groups.

13. *AUPRC:* Used to evaluate binary classification models, especially when the dataset is imbalanced. The AUPRC is computed by plotting Precision against Recall at different thresholds and calculating the area under this curve:

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall})$$

where the integral represents the cumulative area under the precision-recall curve. A higher AUPRC indicates better model performance in distinguishing and separating individuals with a positive versus negative outcome.

14. *Calibration:* the agreement between observed and predicted outcomes. Predicted risks and observed outcomes can be plotted against one another, a diagonal 45 degree line through zero indicates perfect calibration, whilst deviations from this line suggest miscalibration. The calibration plot slope can show if the estimated risks are too homogeneous or too extreme. For example, if 25% of the population drops out and the model predicts that around 25% of students dropped out, then the model is well calibrated on average. The ratio of observed and expected outcomes can be calculated, where a ratio of 1 suggests the calibration is perfect, a ratio above 1 indicates the model predictions are too low and below 1 are too high on average [198].

For simplicity, this chapter uses a 80/10/10 train/validation/test split for the data used to train the TG-CNN models. Previous models using similar data have primarily focused on AUROC, therefore hyperparameters were optimised for best AUROC score on the validation set and test set results are reported.

To optimise the model based on the validation set AUROC value, a random search across

hyperparameters was conducted by sampling the number of epochs [25, 50, 75, 100], learning rate [0.1, 0.01, 0.05, 0.001, 0.005, 0.0001], number of filters [32, 64, 128], filter size [4, 16, 32, 64], number of LSTM hidden cells [16, 32, 64, 128, 256], ℓ_1/ℓ_2 regularisation (Reg) parameter [1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 5e-2, 5e-3, 5e-4], FCL size [128, 256, 512, 1028, 2056], and dropout rate [0.2, 0.3, 0.4, 0.5]. For the 2-stream models, the two streams each had filters of the same size but with different strides. This resulted in 230,400 possible combinations of hyperparameter values, necessitating the use of random search instead of a grid search.

Two baseline (BL) models were also fitted to the dataset for comparison, an BL-LSTM and BL-RNN architecture - a single RNN layer (LSTM or RNN respectively) followed by two FCLs. Five-fold cross-validation was used to optimise the AUROC over the hyperparameter combinations for these baseline models. The BL-LSTM and BL-RNN models were tuned by optimising the learning rate, the number of epochs, the hidden units in the RNNs, and the number of hidden neurons in the FCLs.

4.2.5 Dense Versus Sparse Tensors

Due to the nature of processing these graphs into 3-tensors, most of the indices within the tensors are filled with zeroes. Processing of large tensors, even when the majority is made up of zeroes can be computationally expensive. If more than 90% of the tensor is filled with zeroes, then efficiency is likely to improve with sparse convolutions, otherwise dense matrix multiplication may be more efficient [199]. For this reason, typical dense tensors can be converted to sparse tensors which are more efficient. Figure 4.6 shows how a dense and sparse tensor may be represented as a tensor and using code. Figure 4.6 shows a 2-tensor whereas this project utilises 3-tensors with significantly larger dimensions due to the vast number of clickstream features represented as nodes. Each individual clickstream corresponds to an increase in the size of the tensor in both the x and y axes.

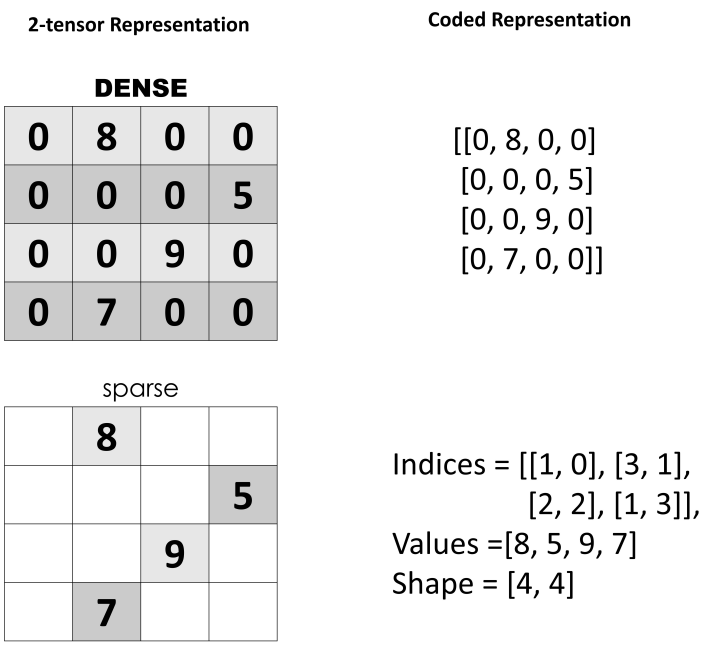


Figure 4.6: Dense and sparse 2-tensors coded.

Performing convolutions on sparse 3-tensors has been shown to improve efficiency, as the convolutional filters do not have to visit spatial locations which have a value of zero [199]. Additionally, smaller filters improve efficiency with sparse CNNs, as hidden layers are often sparse [199].

4.2.6 Speed Comparison

A speed comparison was performed on a dense implementation (which used a standard 3D CNN layer from Keras) and a sparse implementation of the model. Here the model was tested in its most basic form (without regularisation) on a MOOC dataset with the model architecture as shown on Figure 4.5.

The train, validation and test split was 80/10/10 with 5,632 graphs in the training dataset and 704 graphs in each of the validation and test dataset. A Dell Precision 3650 machine, with an 11th generation intel 8 core i7-11700, DDR4 main memory with 64GB of RAM, a NVIDIA GeForce RTX 3090 24GB GPU was used for this speed comparison. Python version 3.9.12 was used with Tensorflow version 2.8.0 and cudnn 8.1.0.77.

To initialise the GPU a TensorFlow constant was loaded as a variable into the memory prior to running the model. The Jupyter Notebook integrated development environment (IDE) was used for speed testing with the kernel restarted prior to each test. A random seed was set so that all the batches were the same across the tests to ensure there were no variations in the

data that could effect the speed. Each test was performed over 50 epochs to obtain an average speed in milliseconds for each sample.

To compare the speed of the dense and sparse 3D CNN layer alone, the start time was initialised when the data is inputted into the model and the end time was allocated as the time when the training data had finishing passing through the 3D CNN layer. The training time was calculated by calculating how long forward and back-propagation took on the training dataset. Inference time was calculated after the model was trained for 1 epoch and this was how long forward propagation took for each sample in the test dataset.

4.3 Results

The TG-CNN models were tested with 1,430 random hyperparameter samples. The best performing hyperparameters and performance metrics for these models are shown in Table 4.1, metrics were averaged over ten runs to show robustness.

BL-LSTM and BL-RNN were each fitted using 5-fold cross-validation with 576 different hyperparameter combinations. The best performing BL-LSTM achieved an AUROC of 0.783, 79.26% accuracy, 0.800 precision, 0.852 recall and an F1-score of 0.824 within 20 epochs, using a learning rate of 0.01, 128 hidden LSTM neurons and FCL sizes of 32 and 16. The best performing BL-RNN achieved an AUROC of 0.778, 78.86% accuracy, 0.801 precision, 0.844 recall, and an F1-score of 0.819 within 20 epochs, with a learning rate of 0.001, 32 hidden RNN neurons, and FCL sizes of 64 and 32.

Table 4.2 shows the best score achieved for each model in the ablation study, compared to existing models in the literature. For MOOC data the best performing variant overall was the TG-CNN with fixed time dilation $\gamma = 1$ (AUROC 0.797), closely followed by the 2-stream version (AUROC 0.796). The average results after 10 re-runs led to the 2-stream model achieving the best performance (Table 4.1). Eliminating the LSTM component led to significantly poorer results (AUROC 0.705).

The results of the speed comparison test can be seen in Table 4.3. The standard deviation was particularly high when using the dense 3D CNN layer as the first epoch took significantly longer than the rest of the epochs due to latency caused by transference of data between the CPU and GPU.

Table 4.1: Hyperparameter values and test set metrics for the best performing variants of the architecture (mean ± standard deviation from 10 runs).

Parameter	Variable γ	1-stream	2-stream	No LSTM	No exp	Elastic Net
Epochs	50	75	52	25	100	100
LR	0.05	0.05	0.0005	0.0001	0.001	0.001
# Filters	64	64	128	64	32	32
Filter Size	16	32	4	4	64	32
RNN	128	64	32	N/A	16	16
L2 Reg	1e-3	1e-2	5e-4	1e-4	5e-2	5e-4
FCL Size	1028	512	2056	512	1024	1024
Dropout	0.5	0.5	0.3	0.5	0.3	0.5
AUROC	0.662±0.08	0.748±0.02	**0.763±0.01**	0.705±0.02	0.710±0.02	0.711±0.02
Accuracy	0.703±0.06	0.771±0.02	**0.775±0.01**	0.583±0.01	0.690±0.02	0.702±0.02
Precision	0.687±0.06	0.764±0.02	**0.773±0.02**	0.581±0.01	0.677±0.02	0.688±0.02
Recall	0.923±0.05	0.883±0.02	0.859±0.04	**1.00±0.00**	0.871±0.02	0.886±0.02
F1-Score	0.782±0.03	**0.817±0.01**	0.811±0.02	0.735±0.01	0.761±0.01	0.774±0.02

Table 4.2: Best area under the receiver operator curve (AUROC) results of user dropout prediction using the ACT MOOC dataset, from our results (left columns) and from the results in the literature (right columns). TG-CNN=temporal graph-based convolutional neural network, BL-LSTM=baseline long short-term memory, BL-RNN=baseline recurrent neural network.

TG-CNN and Baseline Models	**AUROC**		**Literature Models**	**AUROC**
TG-CNN $\gamma = 1$	**0.797**		TGN + MeTA [200]	0.794
TG-CNN 2-stream	0.796		TGN + TNS [201]	0.791
BL-LSTM	0.783		TGN [195]	0.777
BL-RNN	0.779		CoPE [202]	0.762
TG-CNN $\gamma = 4.819$	0.760		JODIE [193]	0.756
TG-CNN with Elastic Net	0.758		TGAT + TNS [201]	0.755
TG-CNN without LSTM	0.750		NPPCTNE [203]	0.745
TG-CNN without the Exponential	0.744		TGAT [200]	0.743

Table 4.3: Speed comparison per sample (in milliseconds) of the dense 3D convolutional neural network (CNN) versus the sparse 3D CNN layered the models (mean ± standard deviation over 50 epochs).

Comparator	Batch Size	Epochs	Dense 3D CNN	Sparse 3D CNN
3D CNN Layer Alone	64	50	0.0256±0.1427	3.8708±0.1096
Training Time	64	50	0.6150±2.6786	5.1469±0.1157
Inference Time	64	50	0.0566±0.0169	3.9671±0.2711
3D CNN Layer Alone	512	50	0.0207±0.1391	0.6067±0.0458
Training Time	512	50	3.1201±21.4529	0.7264±0.0115
Inference Time	512	50	0.0033±0.0074	0.4245±0.0232

Further comparison of the effects of batch size and the layer type on training time speed was carried out with batches of size 32, 64, 128, 512 and 1,024 as seen in Figure 4.7. Smaller batch sizes with the dense 3D CNN layer in the model resulted in faster throughput than the sparse 3D CNN layer, whereas larger batch sizes led to faster performance with a sparse 3D CNN layer. The dense layer speed increases training time almost linearly as batch size increases. Whereas the sparse layer leads to an exponential decay in training speed as the batch sizes increase.

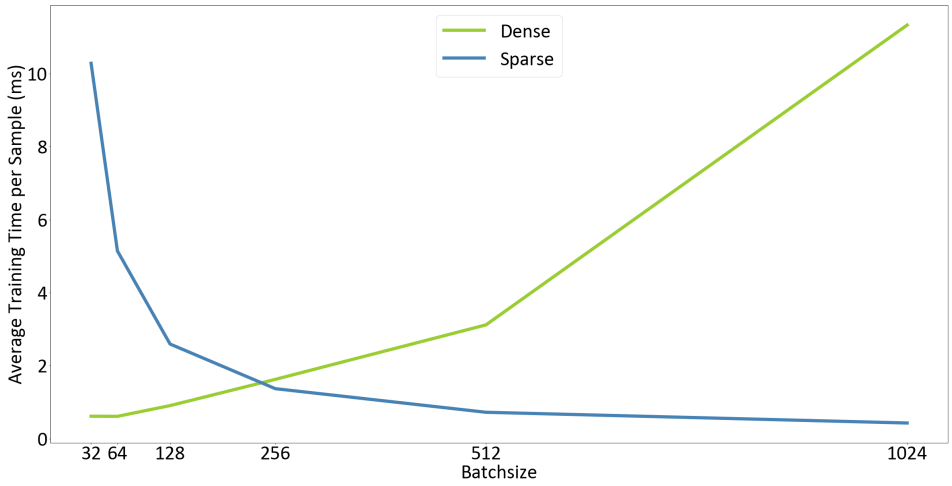


Figure 4.7: Training time per sample for sparse and dense implementation of 3D CNN with varying batch sizes.

4.4 Discussion

Table 4.2 shows the predictive performance of models with the best AUROC score using the ACT MOOC dataset. The TG-CNN model had state-of-the-art performance on this task, whilst being more intuitive and conceptually simpler than some of the other approaches in previous literature.

The ablation study demonstrated that the LSTM layer and the exponential function enabled the model to learn more effectively, this is potentially due to the LSTM layer enabling long-term memory alongside the filters learnt from the CNN. The γ variable converged to an average value of 4.819 in the best performing model, which suggests that actions taken closer together are more important for dropout prediction than actions further apart. When $\gamma = 4.819$ and the elapsed time t is more than 47 seconds $\widehat{t}$ will change to being 0. Having a multi-stream model enables different coarseness of time to be included as features. The second stream with a stride of 2 is a coarser stream which skips more time steps when the CNN filters move across the input data. This means the coarser stream can help find more broader aggregate patterns over time, which might capture long-term trends better. However, this coarse stream might mean the model ignores clickstream actions taken closer together. Having both coarse and fine streams in the model could prevent loss of short-term and long-term patterns, leading to improved performance. The clickstream features may not be highly correlated which led to the elastic net model having poorer performance due to the ℓ_2 regularisation. ℓ_2 alone may also be beneficial as it helps prioritise sparsity and feature selection, which are key to aid in graph interpretability and to reduce computation time. Due to the simplicity of this data, excluding the exponential scaling factor did not lead to memory overflow. However, in larger and more complex datasets this scaling factor may become essential.

Other advantages of the TG-CNN approach include the constant tensor size, allowing for optimisation of the underlying linear algebra operations, and the ability to extract temporal features in parallel using 3D convolutions, as opposed to RNN-based architectures that require sequential processing through time.

The TG-CNN model has interpretability potential, as the 3-tensor structure enables the filters to be extracted back into a intuitive graph structure. This could serve as a visual tool to show which sequences of events and temporal patterns lead to dropout.

4.4.1 Related Work in MOOC Dropout

The IEEE database was searched using the string "MOOC AND predict*". This found 95 papers, 24 of these were full-text analysed based on their title and abstract. Only four used CNNs [204, 205, 206, 207]. Only one paper utilised graphs for node/edge prediction [192]. This differs from our formulation of this problem as a graph-level classification task.

Learner behaviour feature matrices, weighted by importance, have been used alongside CNNs to predict dropout from clickstream data and improved predictive accuracy compared to basic models [191, 204, 205, 206]. Zhang et al 2020. used CNNs alongside Squeeze-and-Excitation Networks (SE-Nets) and a GRU [207]. The GRU enabled maintenance of the time series relationship between the clickstream data and the SE-Net helped with automatic feature extraction. Edmond Meku Fotso et al. found simple RNNs provided better accuracy compared to LSTMs and GRUs [197]. Standard ML algorithms and ensemble methods including SVMs, Logistic Regression (LR), Multi-layered Perceptrons, and DTs have also been applied to this task [191, 194, 208].

The JODIE model (see results in Table 4.2) utilises RNNs to learn and update embeddings that represent individual interactions between users and actions [193]. The actions and users each have their own RNNs to generate separate static and dynamic embeddings. The embeddings dynamically change over time, capturing the temporal aspect in a statically sized graph. These two RNNs are used together for the user embeddings to update the item embedding and vice versa. The JODIE model alters the embeddings significantly after longer periods of time, implicitly assuming that actions taken closer together have smaller impact.

4.4.2 Related Work in Graph Learning

Searching the Web of Science and IEEE databases using the string "Convolution AND (3d OR three$dimension*) AND (time OR temporal) AND graph AND predict* AND network$" returned 18 papers. Of these, there were 5 relevant papers using temporal graph networks [196, 209, 210, 211, 212], though they were all focused on node and edge detection.

At the time of writing (12th May 2022), Kumar et al. had 200 citations of their paper [193]. To observe if any other researchers had used the ACT MOOC dataset processed by Kumar et al. (7,047 users), these 200 papers were screened and 11 papers were found which performed dropout/node prediction tasks.

Four of these utilised RNN components in their model architectures to process time. The others used graph models, all based on a node/ edge classification formulation of the task. Wang et al. used Temporal Graph Networks (TGNs) with dual message passing mechanisms (TGN + MeTA), to augment data and retain semantics for edge-level prediction and node classification [200]. These messaging passing techniques involve memory translation and cross-

level propagation, to adapt the model with temporal and topological features to ignore noise more effectively. This increased the previously obtained AUROC scores by 1.7% [195], with no cost to efficiency and reducing the overfitting that occurs due to noisy data. Other models tweak neighbourhood propagation techniques using temporal information [201, 203]. Zhang et al. use ordinary differential equations and graph networks to observe model changes over time and information propagation [202]. In contrast, the TG-CNN approach makes use of a novel 3-tensor structure, storing the temporal graphs in a sparse and intuitive format, which is easily amenable to feature extraction using convolutions for graph classification.

4.4.3 Limitations

The ACT MOOC dataset used provided clickstream events as numerical labels. Clickstream action descriptions were not provided. Therefore, reasoning for dropout could not be interpreted.

The variant of the model including the time dilation factor γ as a trainable parameter performed the poorest (Table 4.1). The reason behind this is unclear, though the additional complexity modelled by the time dilation will increase the difficulty of the underlying optimisation problem. Due to the random effect of hyperparameter tuning it is possible that a more optimal configuration could be found with further hyperparameter tuning. However, it was not feasible to perform a full grid-search of hyperparameters. Besides the primary objective of this chapter was to produce the methodology rather than produce robust results for this dataset.

Bootstrapping is more robust than splitting data into training and testing groups or cross validation. In future work, bootstrapping should be used to improve result reporting with confidence intervals. Confidence intervals represent a range within which the true result is likely to fall, commonly with 95% confidence if repeated. In frequentist statistics, this interval is constructed under the assumption that if the study were conducted multiple times, 95% of such intervals would contain the true parameter. In contrast, Bayesian credible intervals reflect the probability distribution of the parameter given prior information and observed data [213]. A pragmatic interpretation lies between these perspectives, acknowledging that confidence intervals provide a useful measure of uncertainty even if they do not express a direct probability statement about a single true value. Understanding the assumptions behind each approach ensures appropriate application in statistical inference. Validation of the model on MOOC data could be improved by testing the model on another independent dataset to determine

generalisability.

4.5 Conclusion

This chapter introduced a novel model for the classification of temporal graphs, using student MOOC dropout data to develop and test the method. This approach provides a unique formulation of this problem, compared to previous strategies involving node and edge prediction. This method had AUROC performance improvement in this field compared against the current state-of-the-art models. Methods described in this chapter also benefit from reduced memory utilisation and parallel processing. Proceeding chapters will apply this model to medical data to observe the performance in different scenarios.

Chapter 5

TG-CNNs for Hip Replacement Risk Prediction

5.1 Introduction

The ageing population, coupled with increasing obesity rates, contribute to a rising prevalence of OA and hip replacements, causing pain and diminished quality of life [4, 5, 6, 214, 215]. OA presents a significant challenge for UK healthcare, with MSK issues accounting for up to 1 in 5 of GP consultations [16, 216]. A quarter of all MSK related consultations have been shown to lead to GP advice or no action was taken, 25% provided prescriptions and 16% resulted in a steroid injection [216]. The burden is exacerbated by a growing number of joint replacements, particularly associated to knee and hip OA [38], placing strain on both the NHS and affected individuals [39]. For patients with advanced OA, total joint replacement may be recommended if conservative treatments fail.

MSK conditions impose not only financial burdens but more importantly lead to significant well-being impacts, potentially leading to conditions like depression or obesity [216]. Previous analysis of CPRD data revealed that 32.5% of individuals with hip OA underwent total joint replacement [6]. Between 2003 to 2014, there were 708,311 primary total hip replacements carried out in the UK [38]. Recognising alternative and earlier symptoms and diagnoses associated with hip replacement surgeries may help delay or reduce the need for these invasive procedures.

Predicting hip replacement in advance using primary care data supports efficient resource alloca-

tion and improves patient outcomes. EHR data provides structured and temporal information. However, the irregular time intervals within EHRs present challenges for designing predictive algorithms in healthcare. Using temporal EHR data could assist in accurately forecasting health-related outcomes [66, 111], thereby aiding clinical decision-making and patient care. Additionally, machine learning models can extract valuable insights, such as disease progression, from EHRs [67].

This chapter presents the use of TG-CNNs for individual hip replacement risk prediction one year in advance, using clinical codes and time between primary care visits from patient's EHRs as model inputs. It introduces multiple events per visit/ timestep as parallel nodes, where multiple clinical codes are recorded in a single primary care visit. Within this chapter the TG-CNN model is trained and optimised, investigating which layers, regularisation or other components could be useful to improve predictive performance and minimise overfitting.

5.2 Methodology

5.2.1 Dataset Description and Cohort Analysis

NHS data from ResearchOne was used in this chapter [13]. This data is managed by The Phoenix Partnership (TPP), comprising of clinical and administrative data from 151,565 patients aged 40-75 at the start of the period of analysis, attending healthcare practices in England that use the SystmOne primary care EHR system. This age range was chosen as these patients are likely to present with MSK symptoms. Focusing on those over 40 years old enhances the accuracy and clinical utility of prediction models by targeting the population most at risk for joint replacements (younger patients have less joint wear and usually only need replacements due to rare conditions) and who are most likely to benefit from early identification and intervention. This approach reduces variability and improves generalisability. The data is provided in a de-identified format for patients who: are registered at primary care practices that are opted-in to ResearchOne at practice-level, and who have not opted-out of ResearchOne at patient-level. Patients had their first record of joint pain clinically coded between April 1st, 1999, and March 31st, 2014 [13]. CTV3 clinical codes for around 2,000 different symptoms and conditions were included in this data. Descriptive cohort characteristics are detailed in Table 5.1, prior to removing ineligible records.

Table 5.1: Statistical summary of the electronic health record (EHR) dataset across the analysis period. sd=standard deviation, #=number.

Data collection time period	01/04/1999 - 31/03/2014
Patient age as of 01/04/1999 (mean (sd))	56.67 (9.37)
Total # of patients	151,565
Total # of visits	243,700
Average # of visits per patient	1.827
Total # of unique medical codes	2,353
Average # of medical codes per visit	1.36
Max # of medical codes per visit	10
Max # of medical codes recorded for one patient	534
Average # of visits across the patients	19.99
Max # of visits for one patient	423
Total # deaths	1,039
# of patients who have a hip replacement	12,978

An EHR, in this thesis, is defined as an individual patients p_i record $\mathcal{R}_i = \{r_j | j = 0, ..., N-1\}$, where N is the number of records. For each patient demographic information d_i is kept, including date of birth, sex, and IMD score (a measure of location-based deprivation based on postcodes [217]). Each record r_j contains the reported CTV3 clinical codes c_j alongside the time stamp t. CTV3 clinical codes are clinical codes previously used in the UK. They consist of 5 alphanumeric (upper and lower case) values representing clinical symptoms, diagnoses, procedures and other health related events. CTV3 codes were widely used in primary care before recently being replaced by Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) codes. Patient age at prediction time is used as the age demographic input into our models. IMD score is categorised into quintiles. Including IMD in the model is crucial, as research indicates that individuals with higher IMD scores (indicating lower deprivation) are less likely to undergo hip replacements [6].

5.2.2 Data Extraction

Time windowing: To forecast hip replacements occurring one year in advance, the specific time of the replacement is identified and all available patient records from 1999 up to one year prior to the replacement date are gathered (see Figure 5.1).

Patient inclusion criteria: Patients needed a minimum of two primary care visits within one year prior to a hip replacement (partial or full). Patients were removed if they had no visits in the first 14 years or only had visits in the final year of the analysis period. Clinician-

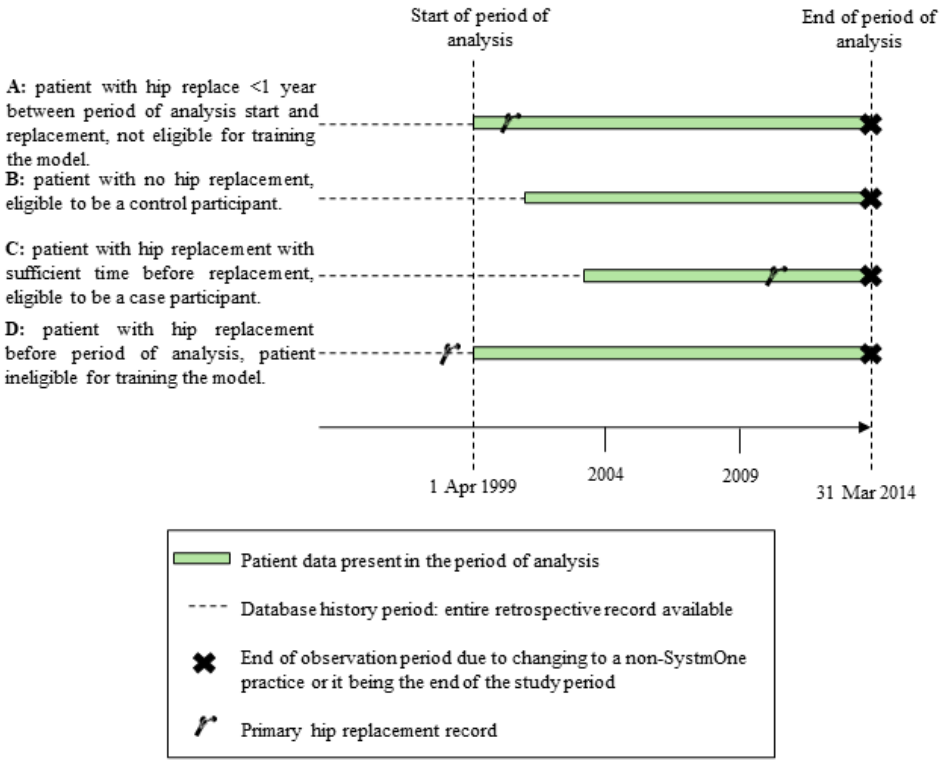


Figure 5.1: Four patient examples (A-D) and their eligibility for the study cohort. Historical records for these patients were included where available. Patient data were included from the start of the analysis period or from their entry into the database if records where present after April 1, 1999 (B and C). Patients were followed until the end of the analysis period or until they changed to a practice not using SystmOne. A primary hip replacement event was considered incident if the first hip replacement was recorded within the analysis period (C).

selected CTV3 clinical codes (n=45) for primary hip replacement were utilised for the outcome, with the incident date defined as the first occurrence of one of these codes. The codes can be found in Appendix B Tables B.1 and B.2. The initial recorded hip replacement instance served as the label in the case patients. To mitigate the influence of age, sex, and IMD on joint replacement, 1:1 exact matching was performed for these variables in the training dataset of each hip replacement patient. This ensured prediction performance was unaffected by differences in these variables. A small fraction (0.007%, n=1,039) of patients in this dataset died, with seven of them having undergone hip replacement. Consequently, complete case analysis was utilised to estimate the marginal effect of hip replacement, assuming no mortality.

5.2.3 Feature Choices

Demographic characteristics, including sex, age, and IMD, were collected alongside all events and medical history within specified time periods. Time between medical events were converted to months, in place of timestamps of data entries. The most frequently used 512 clinical

codes, covering 99.71% of the data, were selected to construct temporal graph-based EHR representations for each patient, ensuring detailed representation while avoiding overfitting and computational memory exhaustion. The first recorded IMD value for each patient was used for consistency, given its stability over time. However, it is important to acknowledge that relying on the first recorded IMD may overlook significant changes in financial and social resources later in life. Older individuals may experience income reduction, loss of family support, or relocation from a more affluent to a less privileged area, which, while not the most common scenario, is particularly relevant for decision-making around OA and hip replacement compared to their earlier IMD score from a previous address. Hospital referral data were not used in this model. In the most frequently used 512 codes there were no predictors synonymous with hip replacement. For the ablation study, medications were appended as six predictors using British National Formulary (BNF) Codes, grouping drugs into opioids (04.07.02), non-opioid analgesics (04.07.01), and NSAIDs (10.01.01), further sub-grouped into acute and repeat prescription types. This addition brought the total number of predictors in the model to 518. Including these predictors in the study nearly tripled the number of recorded events for the included patients in both the training and testing sets. The training dataset without prescriptions had 202,807 rows (clinical codes across all the patients) versus 789,393 rows when prescriptions were included. Similarly, the test dataset had 386,536 rows without prescription data, and 986,326 with.

5.2.4 Temporal Graph Representation of EHRs

In discrete mathematics, a graph $G = (V, E)$ comprises nodes V linked by edges E [137, 138]. This method involves transforming sequences of clinical codes from EHRs into temporal multi-graphs, framing hip replacement risk prognosis as a graph classification problem. Each temporal graph represents an individual and is classified to provide risk probabilities.

Clinical codes were represented by graph nodes (vertex) V, while temporal edges E capture time intervals between code occurrences, measured in months. This is represented in a 3D tensor $G(i, j, k) = t_k$, where $i, j \in \{1, \ldots, n\}$ indicate graph nodes, and t_k is the time elapsed for the kth edge. Each of the 512 most frequently used CTV3 codes maps to one node $v_1, v_2, \ldots, v_{512}$. If a clinical code appears multiple times across visits a new node will be generated for each visit where that code repeats, effectively retaining the sequence of events. Figure 5.2 demonstrates how a snippet of an EHR is converted into a 3-tensor. The maximum number of time steps is

set as $k = 100$, resulting in a 3-tensor input size of $512 \times 512 \times 100$. By representing clinical codes as graphs rather than linear sequences, multiple clinical code recordings per healthcare practice visit can be represented, forming temporal graphs. These graphs can then be fed into the TG-CNN model, enabling pattern detection within clinical code graph structures and temporal features. An overview of using EHR data for hip replacement prediction is provided in Figure 5.3, and the model architecture is discussed in Section 5.2.5.

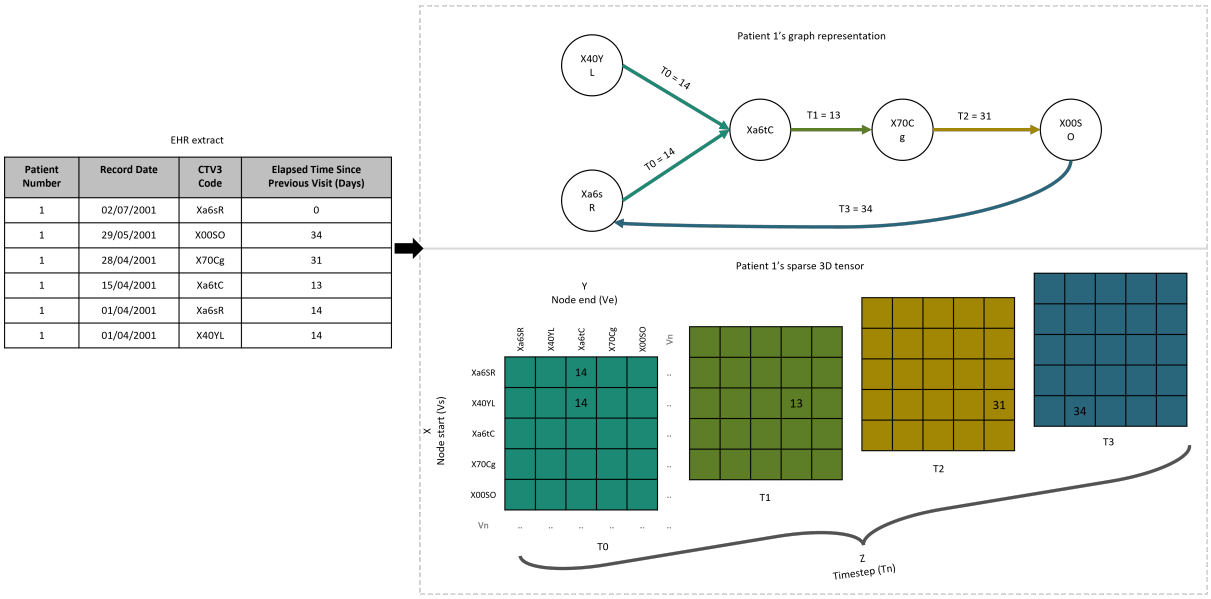


Figure 5.2: Example sequence of five clinical codes being recorded across four time steps (visits). Here there are only 5 clinical codes (nodes), however in reality the unique nodes span to 512 codes which would be difficult to visualise.

5.2.5 Model Architecture

After each patient's EHR is converted to a 3-tensor a neural network architecture is created which applies convolutions over the time axis of the tensor. TensorFlow was employed to craft a custom 3D CNN Keras layer using sparse linear algebra. The temporal graph representation, sized at $512 \times 512 \times 100$, undergoes processing through this layer to extract clinical code sequence and elapsed time patterns (see Figure 5.3). The 3D CNN is similar to standard convolutions used in CNNs for image/video processing, however the filters in this project are much larger in size. A graph of size $n \times n \times k$ (n nodes and k timesteps) is convolved with filters of size $n \times n \times f$, outputting a vector of length $k - f$, assuming a stride length of 1. The CNN output is flattened, followed by batch normalisation for faster convergence, and subsequent layers include a Leaky ReLU, LSTM (captures longer temporal patterns), dropout, dense layers, and concatenation with demographic features (optional). Finally, the model employs binary cross-entropy loss

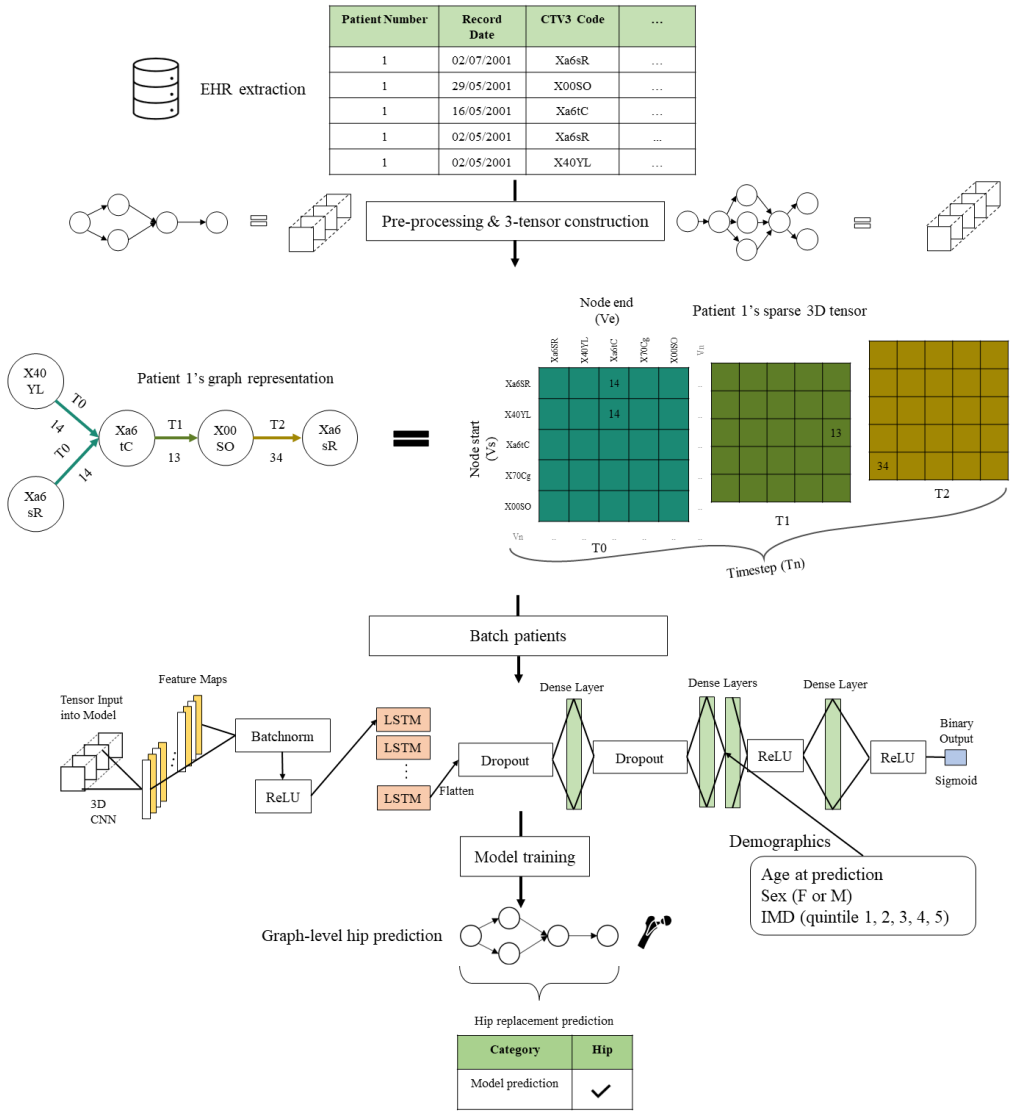


Figure 5.3: Conversion of raw data to model prediction of hip replacement.

with sigmoid for the binary target. Tensorflow 2.8.0, NumPy 1.19.2, Pandas 1.2.4, Scikit-Learn 0.23.1, and CUDA 10.2.89 were used on a desktop with a NVIDIA RTX 3090.

This implementation utilised a learning rate scheduler, multiplying the learning rate by 0.9 with an exponential decay after 1,000 steps. Early stopping was also used with a patience of 5, check-pointing the model when the validation loss decreased and interrupting execution when the model got stuck in a local minima.

Events occurring further in the past may have a diminishing impact on recent events. Therefore, clinicians typically focus on recent health records rather than searching through years' worth of data. Similarly, models predicting patient outcomes prioritise recent events. In this task, the 100 most recent primary care visits per patient were used to manage computational load.

Patient 3-tensors with fewer than 100 visits were front padded, ensuring the most recent visits were always at the end of the 3-tensor. Multi-stream networks were also explored to search over different granularities in time, in the two stream model a stride of one was used as a fine stream and two for a coarse stream. For the two stream models, the two streams each had filters of the same size with only the stride changing.

Events happening in quick succession may indicate a worse health condition compared to those with wider time intervals between them. To capture this, a modified version of elapsed time is stored, measuring the proximity of two events occurring. The graph is transformed using $G(i, j, k) = \exp(-\gamma t)$, where γ is a trainable model parameter. This transformation assigns a score of 1 to simultaneous events and close to 0 for events with a large time gap. Elements indicating no interaction between node pairs at a set time are set to 0.

Filters ideally approximate subgraph pathways, implying that if an individual's EHR is close to a given filter, their disease trajectory should follow a specific path.

ℓ_1 (LASSO) regularisation penalty is employed on the entire filter to prevent overfitting and promote sparsity. This regularisation is applied by adding the ℓ_1 norm to the loss.

ℓ_2 regularisation is a technique used to prevent overfitting by adding a penalty term to the loss function. This penalty is proportional to the sum of the squares of the model's weights. ℓ_2 regularisation discourages large weight values by penalising their square, leading to a smoother and simpler model. By controlling the complexity of the model, ℓ_2 regularisation reduces overfitting to the training data. Unlike ℓ_1 regularisation, which can drive some weights to exactly zero, ℓ_2 regularisation shrinks all weights proportionally but typically keeps them non-zero.

A graph regularisation function is also incorporated, denoted as ℓ_G, which penalises the model with a graph regularisation strength $\lambda_G = 10$ if the filter deviates from a typical graph structure, such as when a node lacks incoming or outgoing connections to other nodes. The penalization factor in this model is the sum of the absolute values of filter weights $|w|$ that are below a predefined threshold (0.1) at specific time steps where no significant connections (weights above the threshold) exist. This factor is computed across all filters and time steps to measure the total deviance from the desired graph structure. The penalty increases as more weights fall below the threshold, especially in cases where no meaningful prior connections are present. This encourages the model to learn graph structures with stronger and more significant connections.

5.2.6 Comparison Models

An ablation study is carried out using 10 variations of the models to assess component benefit as shown in Table 5.2, the full model with the addition of prescriptions was also evaluated.

Table 5.2: Exp = including exponential scaling on the input data. Two streams = trained with an additional stream to integrate a coarse and fine stream of convolutions over the graphs in parallel. w/o=without.

TGCNN Model	**Included**									
Components	γ	Exp	Demo	2^{nd} stream	ℓ_G	ℓ_2	ℓ_1	LSTM	Time	CNN
Full	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
w/o gamma		✓	✓	✓	✓	✓	✓	✓	✓	✓
w/o exp			✓	✓	✓	✓	✓	✓	✓	✓
w/o elapsed time	✓	✓	✓	✓	✓	✓	✓	✓		✓
w/o demographics	✓	✓		✓	✓	✓	✓	✓	✓	✓
w/o two streams	✓	✓	✓		✓	✓	✓	✓	✓	✓
w/o ℓ_G	✓	✓	✓	✓		✓	✓	✓	✓	✓
w/o ℓ_2	✓	✓	✓	✓	✓		✓	✓	✓	✓
w/o ℓ_1	✓	✓	✓	✓	✓	✓		✓	✓	✓
w/o LSTM	✓	✓	✓	✓	✓	✓	✓		✓	✓

When evaluating TG-CNN performance in modelling temporal dependencies within sequential graph data, robust baseline references are crucial. Random Forest and logistic regression were compared with with RNN and LSTM networks. RNNs specialise in processing sequential data by maintaining internal states over time, effectively capturing long-term dependencies within EHR data. However, their performance must be compared against simpler models like Random Forest and logistic regression, which lack explicit mechanisms for modelling such dependencies but offer simplicity and interpretability.

5.2.7 Evaluation Approach

The TRIPOD statement was followed for both the reporting of model development and the prediction models.

A random sample of the 10% of the population was taken which met the inclusion criteria to test the trained model on. This group represented the same proportion of the population that has a hip replacement as the full dataset. From the other 90% every hip replacement patient was matched to a control patient based on sex, IMD, and age at replacement. Oversampling was not performed, each match was a new addition. Within the balanced training dataset 5-fold cross validation was performed when choosing hyperparameters. Training on a balanced

dataset, where each class is equally represented, ensures that the model learns to give equal importance to all classes, preventing it from being biased toward the majority class. To optimise the model based the on the mean of the validation accuracy from cross-validation, a random hyperparameter search was conducted on the Full TG-CNN model 20 times. Hyperparameter values were selected from each of the following lists: learning rate [0.001, 0.005, 0.0001], number of filters [8, 16, 32], filter size [3, 4, 6], number of LSTM hidden cells [16, 32, 64, 128, 256], ℓ_2 regularisation (ℓ_2 reg) parameter [1e-2, 1e-3, 1e-4, 1e-5, 5e-2, 5e-3, 5e-4], FCL size [128, 256, 512, 1028, 2056], and dropout rate [0.5, 0.6, 0.7, 0.8, 0.9]. The model with the best AUROC score from the cross-validation was taken and the same hyperparameters were used across the ablation study models, running each model for a maximum of 12 hours. For the 10% test dataset, half of the dataset was used to recalibrate the model (denoted as Test 1) and the other half (denoted as Test 2) was used to test the recalibrated model (more details to follow in Section 5.2.7). See Figure 5.4 for a visualisation of this data splitting.

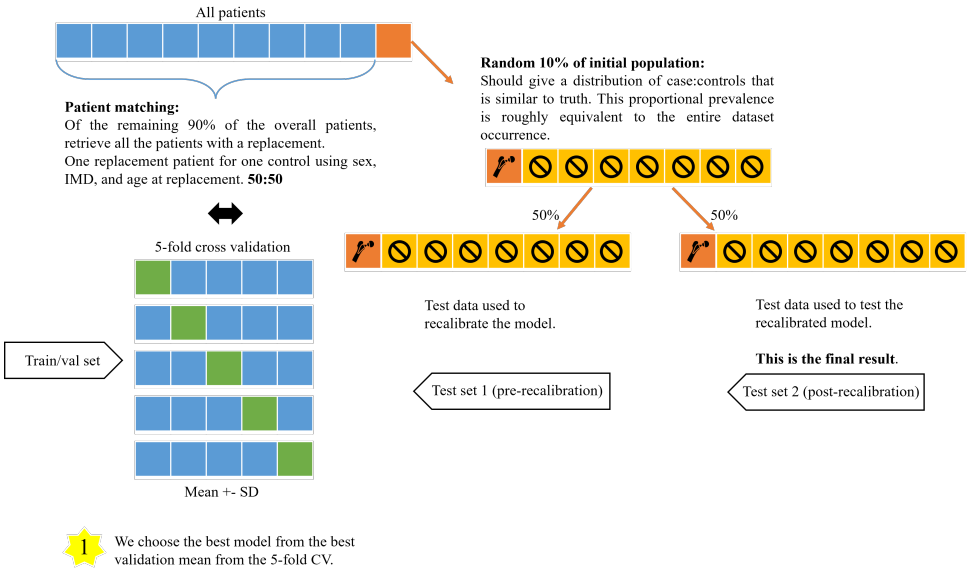


Figure 5.4: How the data was split into training and testing groups.

Calibration was undertaken to assess how close the predicted probabilities reflected the true hip replacement need. Where a well-calibrated model has an average predicted probability in each decile that is close to the average of the actual probability. Calibration is vital to to ensure that the predicted probabilities are accurate for different subgroups based on their risk levels [218]. Discrimination (AUROC and AUPRC) of the models was also assessed.

Clinical data is, more often than not, imbalanced, with more controls than cases. If class imbalance is present prediction errors can occur to favour/ bias the bigger class. Imbalance

can be resolved using techniques such as undersampling or oversampling, weighting approaches, and synthetic minority oversampling (SMOTE) techniques. After training on the balanced data, the model is then recalibrated. Since the model was trained on a balanced dataset, the model predictions were adjusted to enhance alignment with the true incidence rates of the outcomes in the test set. For recalibration, a logistic regression model was created that takes the linear predictors of the original model. The logistic regression model improves calibration by rescaling the linear predictor. The calibration slope can be assessed before and after recalibration to determine efficiency. Once the model is recalibrated on the Test 1 data, the recalibrated model's performance can be validated on a second unseen test set (Test 2). Test 2 data is used to assess model stability, bias, and variability using bootstrapping.

Sub-group analysis (patient stratification) was performed on the best recalibrated TG-CNN model, assessing the models calibration on different demographic sub-groups separately. Sub-groups were created based on sex (female and male), age at prediction (40-60, 60-70 and 70+ year olds) and IMD (quintile).

5.3 Results

Table 5.4 displays the characteristics of the cohort included in model training and testing. The Body Mass Index (BMI) statistics in this table were derived from the last recorded BMI measurement of each patient. The last-recorded BMI was used simply for a cohort characteristic summary in Table 5.4, the TG-CNN model takes account of BMI changes over time as nodes. The last-recorded BMI provided in this Table may not accurately represent the historical BMI that contributed to the development of OA and the need for hip replacement. As disease or disability progresses, loss of skeletal muscle mass can lower BMI, potentially masking earlier periods of high adiposity. Earlier, higher BMI measurements may provide a more accurate reflection of the excess fat that influenced OA progression over time.

A Chi-squared analysis was conducted to compare the population of Test 2 to the National Joint Registry 12th Annual Report 2015, to ensure similar demographic distributions [38]. Results in Table 5.3 show no significant difference in age, sex, and IMD between our Test 2 set and the National Joint Registry. However, there was a significant difference observed in weight category distribution between the datasets.

Table 5.3: Chi-squared analysis comparing Test Set 2 dataset to National joint registry (NJR) 2015 population. k = 2, df = 1, alpha = 0.05, and Chi-square value = 3.841. If $\frac{(O-E)^2}{E} > 3.841$, the values are significantly different. Where O = *Observed*, E = *Expected*. IMD=index of multiple deprivation.

	Test Set 2 (N (%))	**NJR (N (%))**	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Female	3512 (60.6)	21304 (60.0)	0.60	0.36	0.01
Age (median)	66	69	-3.00	9.00	0.13
IMD 1	1565 (21.3)	5406 (15.2)	6.10	37.22	2.45
IMD 2	1554 (21.1)	6532 (18.4)	2.79	7.78	0.42
IMD 3	1373 (18.7)	8064 (22.7)	-3.98	15.80	0.70
IMD 4	1324 (18.0)	8084 (22.7)	-4.70	22.07	0.97
IMD 5	1537 (20.9)	7520 (21.1)	-0.22	0.05	0.00
Obese	2108 (31.7)	15270 (55.8)	-24.09	580.45	10.41
Overweight	2674 (40.2)	9407 (34.4)	5.82	33.86	0.99
Normal	1773 (26.6)	2634 (9.6)	17.02	289.53	30.10
Underweight	102 (1.5)	75 (0.3)	1.26	1.58	5.78

Table 5.5 displays the distribution of hip replacements per deprivation quintile, indicating slightly higher rates in more deprived areas (IMD 1-3) compared to less deprived regions.

Upon running the trained model on the test data, an inverted calibration curve was observed (Figure 5.5). This suggests that the model was underpredicting the likelihood of patients needing a hip replacement and that the model's predicted probabilities were too low and not aligned with actual observed outcomes. However, recalibrating the model resulted in a closer fit to the optimal line and a more accurate probability distribution. To understand this phenomenon, several evaluations were conducted. The model was trained on an imbalanced dataset, where each case patient was matched with two control patients, yet the probability distribution remained skewed towards case patients. Further testing involved reserving 20% of unmatched control patients for testing, resulting in similar probability skewing. This phenomenon persisted during training of baseline models, indicating it was data-based rather than model-based. Holding back a portion of matched training data as test data yielded good results even prior to calibration. These tests suggest that the occurrence is due to the model being well-trained on matched patients. The patient trajectories will be explored further in Chapter 7 the explainability chapter.

The full TG-CNN model with the best validation score from 5-fold cross-validation had the following hyperparameters: learning rate = 0.0001, filters = 32, filter size = 6, LSTM neurons = 128, FCL size = 128, dropout value = 0.6, regularisation strength = 0.0005, and graph regularisation strength = 0.1.

Table 5.4: Hip replacement prediction model cohort characteristics. BMI=body mass index, IMD=index of multiple deprivation, std=standard deviation.

Characteristic	**Overall (N=33,080)**	**Train (N=18,374)**	**Test 1 (N=7,353)**	**Test 2 (N=7,353)**
Sex: female	20,501 (62.0)	11,626(63.3)	4,422(60.2)	4,453(60.6)
Hip replacements	10,214 (30.9)	9,187 (50)	515 (7.0)	515 (7.0)
Age, y				
mean(std)	70.4 (9.2)	72.5(8.5)	67.9(9.3)	67.6(9.3)
median(min,max)	70 (53,89)	73(53,89)	67(53,89)	66(53,89)
mode	66	76	66	66
Age at replace, y				
mean(std)	69.2 (8.5)	69.2(8.53)	69.6(8.6)	69.4(8.8)
median(min,max)	70 (43,90)	70(43,90)	70(48,89)	70(48,88)
mode	77	77	78	77
BMI				
Severely obese	1,204 (3.6)	652(3.5)	267(3.6)	285(3.9)
Obese	9,946 (30.1)	5,803(31.6)	2,035(27.7)	2,108(28.7)
Overweight	12,089 (36.5)	6,721(36.6)	2,694(36.6)	2,674(36.4)
Healthy weight	7,793 (23.6)	4,304(23.4)	1,716(23.3)	1,773(24.1)
Underweight	491 (1.5)	287(1.6)	102(1.4)	102(1.4)
Missing	1,557(4.7)	607(3.3)	539(7.3)	411(5.6)
IMD, quintile				
1 (most deprived)	8,370 (25.3)	4,996(27.2)	1,809(24.6)	1,565(21.3)
2	7,859 (23.8)	4,640(25.3)	1,665(22.6)	1,554(21.1)
3	6,990 (21.1)	3,982(21.7)	1,635(22.2)	1,373(18.7)
4	5,195 (15.7)	2,648(14.4)	1,223(16.6)	1,324(18.0)
5 (least deprived)	4,666 (14.1)	2,108(11.5)	1,021(13.9)	1,537(20.9)
Events Recorded				
mean(std)	17.3 (20.0)	10.6 (13.5)	24.7(22.8)	26.7(23.5)
median(min,max)	10.0 (2,224)	6.0 (2,161)	18.0 (2,224)	20.0(2,182)
mode	3	3	8	7

Table 5.6 displays AUROC and calibration slope results from 5-fold cross-validation on the TG-CNN model. The w/o ℓ_1 model was selected to illustrate before and after recalibration calibration plots (Figures 5.5 and 5.6), as this model demonstrates the best calibration (with the validation C-slope value closest to 1). The TG-CNN w/o exponential model outperformed the others in terms of AUROC on the balanced dataset but showed poor recalibration, suggesting overfitting. Table 5.7 and Table 5.8 present AUROC and AUPRC results using the Test 2 (unseen) data on the TG-CNN and baseline models, respectively.

The most important features, according to the best Random Forest model are provided in Figure 5.7. Additionally, Figure 5.8 displays the top 10 largest odds ratios from the logistic regression model, indicating variables that act as risk multipliers. As anticipated, both models seem to recognise that clinical codes associated to hip pain and OA are crucial for predicting

Table 5.5: Hip replacement occurrence in each index of multiple deprivation (IMD) group N(%).

	Overall	**Train**	**Test 1 (Pre-recalibration)**	**Test 2 (Post-recalibration)**
IMD 1	2,750(32.9)	2,498(50.0)	142(7.9)	110(7.0)
IMD 2	2,549(32.4)	2,320(50.0)	102(6.1)	127(8.2)
IMD 3	2,248(32.2)	1,991(50.0)	133(8.1)	124(9.0)
IMD 4	1,498(28.8)	1,324(50.0)	90(7.4)	84(6.3)
IMD 5	1,172(25.1)	1,054(50.0)	48(4.7)	70(4.6)

Table 5.6: Area under the receiver operator curve (AUROC) and C-slope (mean (sd)) results for the models on the training set.

TGCNN Model	**Train AUROC**	**Validation AUROC**	**Train C-slope**	**Validation C-slope**
w/o exp	0.969 (0.003)	0.936 (0.009)	0.999 (0.011)	0.816 (0.059)
w/o ℓ_1	0.937 (0.007)	0.882 (0.014)	1.044 (0.016)	1.003 (0.086)
w/o LSTM	0.920 (0.004)	0.859 (0.017)	0.840 (0.146)	0.891 (0.067)
w/o ℓ_G	0.973 (0.010)	0.858 (0.009)	1.043 (0.051)	0.613 (0.061)
w/o demographics	0.959 (0.038)	0.852 (0.009)	1.058 (0.021)	0.677 (0.113)
Full	0.966 (0.016)	0.850 (0.002)	1.021 (0.029)	0.685 (0.132)
w/o elapsed time	0.966 (0.016)	0.844 (0.009)	0.100 (0.054)	0.585 (0.082)
w/o ℓ_2	0.971 (0.007)	0.840 (0.020)	1.018 (0.028)	0.603 (0.123)
w/o two streams	0.521 (0.003)	0.519 (0.009)	0.867 (0.254)	0.863 (0.595)
w/o γ	0.518 (0.002)	0.516 (0.005)	1.074 (0.153)	0.963 (0.563)
with prescriptions	0.516 (0.003)	0.517 (0.008)	0.820 (0.191)	0.675 (0.446)

Table 5.7: AUROC and AUPRC results for the recalibrated TG-CNN models on Test 2 data.

TGCNN Model	**AUROC mean (sd)**	**AUROC (95% CI)**	**AUPRC mean (sd)**	**AUPRC (95% CI)**
w/o ℓ_2	**0.724 (0.014)**	(0.715,0.733)	0.185 (0.013)	(0.160, 0.209)
w/o LSTM	0.717 (0.007)	(0.713,0.721)	0.160 (0.011)	(0.139, 0.182)
w/o ℓ_1	0.716 (0.018)	(0.705,0.728)	**0.220 (0.029)**	(0.163, 0.277)
w/o ℓ_G	0.699 (0.015)	(0.689, 0.708)	0.165 (0.008)	(0.149, 0.182)
w/o demographics	0.696 (0.021)	(0.683, 0.709)	0.151 (0.018)	(0.116, 0.187)
Full	0.693 (0.015)	(0.684, 0.702)	0.165 (0.011)	(0.143, 0.188)
with prescriptions	0.684 (0.008)	(0.680, 0.689)	0.116 (0.004)	(0.108, 0.125)
w/o elapsed time	0.680 (0.013)	(0.672, 0.688)	0.150 (0.013)	(0.124, 0.176)
w/o exp	0.634 (0.019)	(0.622, 0.646)	0.169 (0.011)	(0.148, 0.190)
w/o γ	0.589 (0.009)	(0.584, 0.595)	0.083 (0.005)	(0.073, 0.092)
w/o two streams	0.570 (0.009)	(0.564, 0.575)	0.080 (0.006)	(0.069, 0.091)

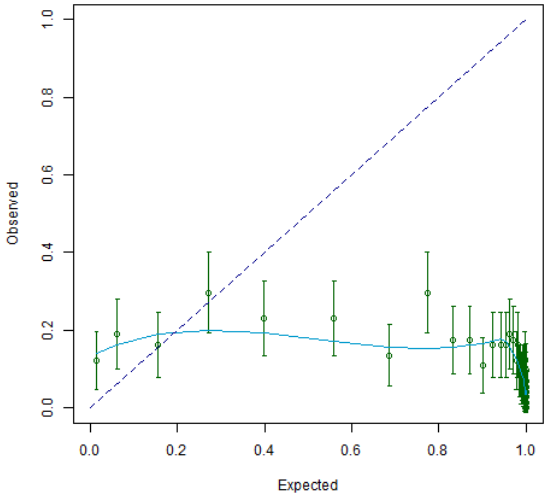


Figure 5.5: Pre-recalibration calibration curve on Test 1 data. Each green bar represents 1% of the patients.

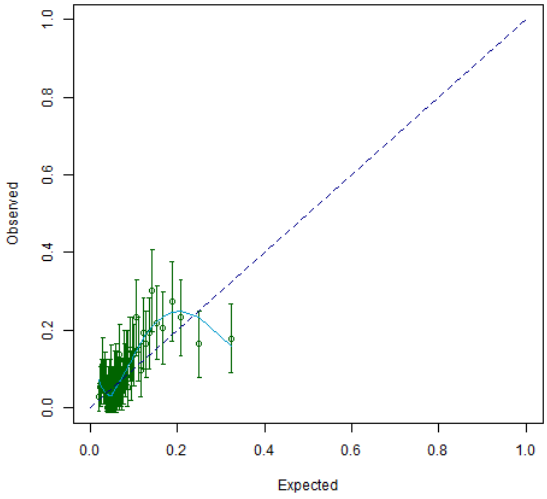


Figure 5.6: Post-recalibration calibration on the Test 2 data. Each green bar represents 1% of the patients.

Table 5.8: Results for the recalibrated baseline models on the unseen Test 2 set. RNN=recurrent neural network, LSTM=long short-term memory, LR=Logistic regression model, RF=Random Forest.

Model	**AUROC mean(std)**	**AUROC (95% CI)**	**AUPRC mean(std)**	**AUPRC (95% CI)**
RNN	**0.698 (0.014)**	(0.689, 0.706)	0.135 (0.010)	(0.116, 0.155)
LSTM	0.651 (0.011)	(0.644, 0.658)	0.120 (0.008)	(0.105, 0.136)
LR demo only	0.580 (0.015)	(0.571, 0.590)	0.080 (0.004)	(0.073, 0.087)
RF demo only	0.575 (0.009)	(0.570, 0.581)	0.080 (0.004)	(0.072, 0.088)
LR CTV3 only	0.529 (0.018)	(0.519, 0.540)	0.090 (0.010)	(0.070, 0.110)
LR demo and CTV3	0.529 (0.014)	(0.520, 0.537)	0.088 (0.006)	(0.076, 0.100)
RF CTV3 only	0.528 (0.011)	(0.521, 0.535)	0.086 (0.003)	(0.080, 0.091)
RF demo and CTV3	0.518 (0.014)	(0.510, 0.527)	0.083 (0.006)	(0.071, 0.096)

hip replacement.

Subgroup analysis (patient stratification) was conducted on the recalibrated TG-CNN w/o ℓ_1 model. See the Appendix for calibration curves (Figures B.1 - B.10) which demonstrate equitable performance across demographic subgroups. There seemed to be slight underprediction in the female population compared to males, but overall, predicted risks were reasonable in a large proportion of the population. Younger populations tended to experience more overprediction, while individuals aged 60-70 exhibited the best calibration; however, those over 70 are notably underpredicted. Regarding deprivation levels, IMD 1, 2, and 4 demonstrated the best calibration, while IMD 5 overpredicted hip replacement and IMD 3 underpredicted, indicating

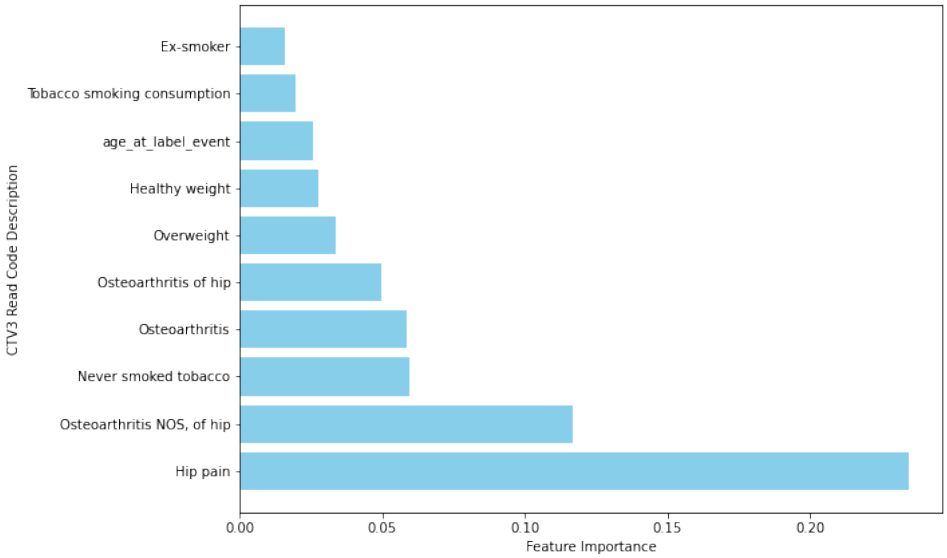


Figure 5.7: The 10 most important features from the best Random Forest model. NOS=Not otherwise specified.

no bias towards more or less deprived areas.

Figure 5.9 shows the converged upon γ values for each of the trained models. These values demonstrate how proximity of visits effects each of the model differently.

5.4 Discussion

Finding patterns in MSK outcomes in the case of total joint replacement follow-ups, could be vital to improve post-operative outcomes. A recent study showed that joint replacement routine follow-up appointments are unnecessary 1-10 years after hip or knee replacement [219]. However, the association between joint pain and health trajectories leading up to joint replacement is currently not well known.

Yu et al. investigated predicting the risk of hip replacement among patients with OA or hip pain in the UK using CPRD data. They developed cumulative incidence equations to forecast the 10-year probability of hip replacement, narrowing down the model to 20 predictors for hip replacement. They achieved an AUROC of 0.72 and a C-slope of 1.00. This study marked the first clinical risk prediction model for hip replacement in patients newly presenting with hip pain/OA in primary care. Hip injury strongly indicated future hip replacement risk [14]. Other models, such as those by [220] and [221], employ radiographic images for prediction. Utilising TG-CNNs and EHR data from primary care services to predict hip replacement risk

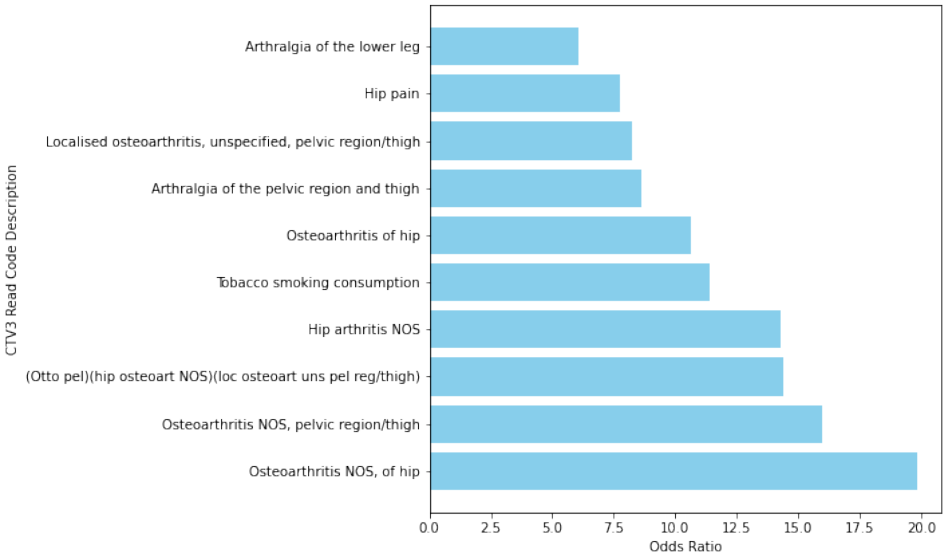


Figure 5.8: Top 10 most influential predictors from the best Logistic Regression model.

could reveal novel patterns associated with the procedure, thus enhancing predictive accuracy. This approach could facilitate early intervention or triage for individuals showing potential hip replacement indicators.

In a study on deep learning methods for handling irregularly sampled medical time series data [222], linear regression, Random Forest, and SVMs showed predictive capabilities for health outcomes but struggled with integrating irregular time data effectively. Conversely, T-LSTM and DATA-GRU models were suitable for accommodating irregularly sampled data, with DATA-GRU featuring implicit attention mechanisms that enhance interpretability. The TG-CNN model generates informative features, evaluates events within their clinical context to reduce reporting bias, and incorporates elapsed time between events to enhance predictive accuracy.

As Chapter 4 demonstrated the capability of predicting student dropout using MOOC data, there is a key distinction in the medical domain when using EHRs. Click actions can only occur sequentially, whilst multiple clinical codes can be recorded in a single visit, leading to parallelism within graph networks.

The Random Forest and logistic regression models, though interpretable with top predictors like hip pain, hip OA, OA, and tobacco smoking, showed poor AUROC and AUPRC. The Random Forest model emphasised OA, smoking, and weight. Using demographics alone improved performance slightly compared to using clinical codes in the recalibrated Random Forest and logistic regression models. While these models performed competitively on the balanced train-

ing set, their recalibrated application to Test 2 data led to poorer AUPRC results. Excluding demographics in the TG-CNN model resulted in overfitting, indicating that including predictors beyond basic demographic information may be beneficial. There is a known association between sex, IMD, and age regarding the likelihood of needing a hip replacement if someone has OA, justifying their inclusion.

The LSTM and RNN baseline models might have performed poorly due to time decay on events occurring further in the past, weakening relationships with longer time intervals. Reversing the input vector (such that more recent events were at the front instead of back of the tensor) prevented convergence, indicating that the model may learn best from a mix of recent and historical events. With simple RNNs, elapsed time cannot be included, so the events were fed into the model as a sequence, leading to the loss of some temporal features. Including an LSTM in the TG-CNN model could be beneficial for retaining distant clinical codes in memory rather than focusing solely on recent ones.

Using an exponential function for elapsed time allows rescaling to avoid extreme values in neural networks and mitigate potential under/ overflow issues with half-precision arithmetic. The negative exponential ensures that actions taken quickly in succession have values close to 1, while those with greater temporal gaps are closer to 0. The model without the exponential component achieved the highest validation AUROC with minimal difference from the training set, suggesting no overfitting occurred. However, despite its success on balanced data during training, it performed poorly during recalibration, leading to significant underprediction for some patients. This suggests that excluding the exponential component reduces generalisability to unbalanced datasets. It also implies that the time step (axis j of the 3D tensor representing the temporal graph) requires the exponential component to recognise whether events are more distant or recent.

Including the trainable γ parameter allows the model to shift the exponentially scaled value of the elapsed time to optimise the model further. The γ value was expected to increase the difference between the more recent and historical codes, such that older codes are scaled to be smaller than more recent events which are inflated. However, the TG-CNN model without the variable γ would not converge which suggested that the inflation/ shrinkage to the elapsed time is too strong on the model and that past events may be more influential to the outcome prediction than anticipated. Figure 5.9 shows that γ ranges between -0.1 and 0.15, this suggests very minor

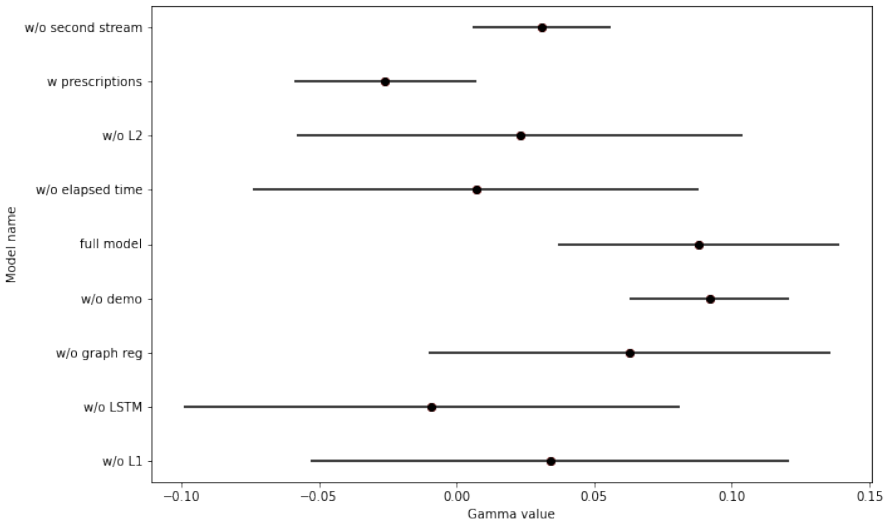


Figure 5.9: Average (and standard deviation) trained gamma (γ) value from cross-validation from each of the TG-CNN models that incorporate γ.

scaling of the elapsed time is beneficial to the model, and that the more regularly positive γ's suggest that perhaps events happening closer in time are more important to prediction than events further in the past. The LSTM helps preserve the memory of events happening further in the past, so by not having the LSTM component in the model the γ value decreased to compensate and give less importance to recent events as the prior events. This could be because events far in the past are not well stored by the 3D CNN layer alone. Similarly, the prescriptions model had an average negative γ, perhaps as larger gaps between codes are more important due to the more recent codes being prescription related making it difficult to distinguish between classes.

Incorporating prescriptions drastically increased the number of eligible records for the TG-CNN model. Including prescriptions in the model may result in poor validation AUROC outcomes because the graphs become inundated with prescriptions rather than diagnostic clinical codes. Moreover, additional features could increase model complexity by expanding dimensionality, potentially prolonging convergence time and yielding poorer results.

Elastic Net might not perform as effectively as ℓ_2 regularisation alone, especially when handling highly correlated features, as indicated by the good performance of the TG-CNN model without ℓ_1 regularisation. ℓ_1 regularisation tends to select only one feature from a group of correlated features, effectively disregarding the rest which is good for sparsity, as it can shrink values to 0. In contrast, ℓ_2 regularisation penalises all feature coefficients, distributing the weight more

evenly among correlated features without shrinkage to 0. Our custom graph regularisation function ℓ_G seems to effectively prevent overfitting.

The model without a second stream would not converge, this shows that having a long and a short stream is useful for seeing both large and small features. Having two different stride lengths capture features at different scales. A fine branch with smaller strides might capture finer details over each visit, while a coarse branch with larger strides captures broader patterns and structures. Combining these branches allows the model to learn features at multiple scales simultaneously, enhancing its ability to understand complex patterns in the data.

The TG-CNN model without elapsed time overfitted and exhibited a poor C-slope value, highlighting the significance of incorporating irregular elapsed time between clinical code recordings as a crucial feature to enhance predictive performance.

The TG-CNN model utilises EHR data from primary care, allowing for the identification of individuals at higher risk of future hip replacement. With this model, targeted preventative care could be applied to slow progression, severity, or pain. Future research should investigate clinical outcomes and potential interventions in primary care based on this model.

The TG-CNN model serves as an alternative to the current EHR-based hip prediction model [14], enabling the infusion of temporality and trajectory. In future chapters explainability is incorporated into this approach using the filters from the CNN layer, to enable further trust in the model and explain why certain predictions may have occurred [223]. The next chapter explores the reconfiguration of the TG-CNN model to determine the risk of hip or knee replacement 5 years in advance, allowing healthcare professionals more time to apply interventions and planning.

5.4.1 Limitations

When models predict health data outcomes it is important that they are validated appropriately. Whilst metrics such as sensitivity, accuracy and AUROC are valuable they do not fully describe the net impact for patients if the model predicts incorrectly. They also do not allow comparison of how the model compares to real-world prediction, for example comparing against the accuracy of a clinician's predictions.

Decision curve analysis can also be informative to calculate the benefit of the model depend-

ing on risk thresholds for clinical practice utilisation. What would the result be of a model incorrectly predicting an outcome for a patient, could it worsen a patient's health significantly? What is the result of giving no intervention versus an unnecessary intervention? These can be weighted to show which has a higher impact (benefit or harm) on patient health. Additionally as some clinicians and patients have different perceptions of life impact, the thresholds can be tailored to suit a wider range of audiences.

Ideally external validation should be carried out, this should involve a completely unseen dataset, unrelated to the dataset that model training was performed on. Alternatively, chronological splitting (testing the model at different stages of history) or splitting by health centre should be carried out to ensure generalisability across different time, locations, and settings. External validation was not feasible in this thesis as only one healthcare dataset with time stamps between visits was available during this project, significantly more time would be required to apply and get access to a secondary appropriate dataset.

Studies have indicated that knee OA may be diagnosed narratively (i.e., via free-text notes written by clinicians) an average of three years before being codified in EHRs. Approximately one-third of patients receive a narrative diagnosis of knee OA before a codified diagnosis appears in their records [215]. The prevalence of knee OA nearly doubles when free-text notes are included alongside clinical codes. From 2008 to 2019, the prevalence of narratively diagnosed knee OA without a corresponding clinical code record increased from 2.92% to 5.60% [215].

Furthermore, GPs may lack the time to code diagnoses or symptoms during consultations, or they may select incorrect codes, potentially leading to misclassification. Also events recorded on a specific day in EHRs may not have actually occurred on that day, leading to inaccuracies in the time between EHRs. To identify patients with OA earlier and more accurately, incorporating narrative data into predictive models might be worth considering, rather than relying solely on clinical codes.

Although multi-modal models that include imaging could be explored, healthcare constraints on cost and time often make it more practical to work with faster-to-train models, which are easier to recalibrate to prevent temporal drift. By utilising data readily available from EHRs, it is believed that implementation into routine primary care would become more feasible. Additionally, while full imaging data could provide rich model predictors, metrics derived from images (which are often reported in radiology assessments), may already capture highly predictive features.

Using structured radiology data, rather than raw image inputs, could streamline model training processes by reducing computational burden whilst maintaining predictive power [224].

Left censoring issues, whereby clinical codes are only available from a specific date onward, may result in joint replacements that occurred prior to the analysis period going unrecognised. Conversely, right censoring issues indicate that a hip replacement could take place shortly after the analysis period, meaning that such patients may not be accurately classified as controls.

It should also be recognised that, even when a hip replacement is offered, some patients may choose not to proceed. Consequently, health events could indicate the need for a replacement, yet some patients may not undergo the procedure, introducing potential errors into the model.

The CTV3 clinical coding system is no longer being used, however, the TG-CNN methodology described in this thesis could easily be transferred to other coding systems.

This dataset does not contain information on ethnicity, which is an important feature relevant to healthcare outcomes in patients. Additionally, due to patient re-identification risks IMD score is the only feature provided in this dataset to score socioeconomic status of patients, however this is based on postcodes rather than individual-level deprivation so may not be an accurate representation. The dataset does not include limb sidedness/ laterality so it cannot be specified which leg has had the replacement. This means that a 'primary replacement' may be someone's second primary replacement (but on a different leg), the first record of primary replacement is used to try and mitigate for this. The medical reasons for replacements or whether the replacement was elective or not was not available, but this could give insights if a replacement was due to trauma or other health conditions. A hip replacement may also be the only way to reduce pain and improve mobility, but the existence of other co-morbidities may be so severe as to potentially have higher surgical risk, such that the patient may not wish to proceed. Additionally, if the patient has caring responsibilities, they might not wish to have the surgery as they do not believe they will be able to have recovery time whilst caring.

5.5 Conclusion

The TG-CNN model can aid in clinical decision-making by targeting individuals at high risk of future hip replacement for intensive non-surgical management or active monitoring, enabling the application of preventive treatments and care. Implementing this model clinically may po-

tentially reduce patients' time in pain, enhance quality of life, and improve healthcare efficiency and resource allocation.

Chapter 6

Hip and Knee Replacement Risk Prediction

6.1 Introduction

OA is associated with decreased physical activity, severe pain, and increased morbidity [26]. 32.5% of those with hip OA and 16.7% with knee OA eventually require joint replacement [6].

In 2014, the UK National Joint Registry recorded 83,125 hip replacements, with each procedure costing the UK NHS an average of £5,280, resulting in an annual expenditure of £438.9 million for primary hip replacements alone [38, 50]. This cost is significant but essential to improve patient quality of life [225]. In 2009, OA was the fourth most common reason for hospitalisation [37]. 25% of adults are expected to develop symptomatic OA [37]. Pressures on hospitals will rise and with more people in pain as OA prevalence increases with ageing populations [38, 39]. From 2003 to 2014, 708,311 primary total hip replacements and 772,818 primary knee replacements were performed in the UK, with OA accounting for 93% and 96% of these procedures, respectively [38]. The risk of knee replacement at time of knee OA diagnosis increases up until age 62 then decreases, whilst hip replacement at time of hip OA diagnosis decreases linearly the older the patient is [47]. Due to growing OA cases, unlicensed prescriptions are increasingly issued as clinicians seek alternative painkillers to alleviate patient suffering [48].

Clinical prediction models can estimate an individual's risk of future medical events, such as diseases or drug success, prior to any physical tests. ML enables detection of health patterns

utilising machine memory and processing power. Predicting the need for future hip or knee replacements is valuable for resource planning, triaging, clinician decision-making, and early intervention. For example, exercise therapy has been shown to reduce hip replacements by 44% [8], and early interventions like physiotherapy can improve patient quality of life and lower surgery rates.

In the previous chapter, the TG-CNN methodology was introduced for the prediction of hip replacement risk one-year in advance using EHR CTV3 codes and demographics from ResearchOne data. In this chapter, four separate models are created for hip and knee prediction, with replacement risk predicted one and five years in advance, with more distant prediction potentially allowing for better resource planning.

Predicting hip and knee replacements at different time horizons provides significant benefits for both patients and healthcare systems. A five-year prediction window allows for early interventions that may delay or even prevent the need for surgery. Patients identified as high-risk can be offered conservative treatments such as physiotherapy, weight management, and lifestyle modifications, potentially improving joint health and slowing disease progression. Additionally, long-term predictions enable healthcare systems to better allocate resources and plan for future surgical demand. In contrast, a one-year prediction timeframe is particularly valuable for short-term surgical planning, allowing hospitals to manage waiting lists, allocate resources efficiently, and ensure patients receive timely care. Patients can also be better prepared for surgery through prehabilitation, which can optimise surgical outcomes and recovery.

Extending the model to knee replacements is particularly important, as knee OA is more prevalent than hip OA and accounts for a significant proportion of joint replacement surgeries. Knee replacements also have different recovery trajectories and rehabilitation needs, making tailored prediction models essential. Identifying knee replacement candidates early can help mitigate mobility loss, pain progression, and reduced quality of life, particularly given the functional demands of the knee joint in daily activities. By incorporating both hip and knee replacements across short- and long-term timeframes, this approach enhances patient care, improves healthcare efficiency, and optimises overall outcomes.

The IMD score was included in these models as an indicator of sociodemographic status, based on research showing that individuals with higher IMD scores (indicating less deprivation) were less likely to have hip replacements [6]. This suggests a lower incidence among the more afflu-

ent populations. However, this does not necessarily imply equity, as multiple factors influence surgical access, including health-seeking behaviours, clinical decision-making, and availability of healthcare resources. Individuals in higher deprivation groups tend to have greater musculoskeletal disease burden and higher need for joint replacements [226], yet they may experience barriers to accessing surgery, such as longer wait times, lower referral rates, and financial constraints affecting post-surgical recovery [227]. Consequently, while the incidence of hip replacements may be higher in certain lower socioeconomic groups, this does not inherently indicate equitable access.

Additionally, age is a critical predictor, with knee replacement risk peaking in 60-70 year olds, and hip replacements peaking at 55 [38, 47]. Sex is also significant, as women were found to be more likely to have joint replacements, comprising 60% of hip replacement surgeries and 57% of knee replacements [38].

The aim of this chapter is to predict hip or knee replacement risk one and five-years in advance, using CTV3 codes constructed as temporal graphs for prediction using the TG-CNN model. The results from this model are compared to the current state-of-the-art in the literature, and subgroup analysis is carried out to see the effect on model performance in different patient groups.

6.2 Methodology

The methodology builds upon the approaches outlined in Chapters 4 and 5, applying the hip replacement risk model to knee replacements and extending the prediction time frames to both one year and five years in advance.

Sequences of CTV3 codes from EHRs were transformed into temporal multigraphs, framing hip and knee replacement risk prognosis as a graph classification problem. Each temporal graph, representing an individual, was classified to provide risk probabilities.

CTV3 codes were used as graph nodes V, while temporal edges E captured time intervals between code occurrences, measured in months, and represented in a 3D tensor $G(i, j, k) = t_k$, where $i, j \in \{1, \ldots, n\}$ indicate graph nodes, and t_k is the time elapsed for the k^{th} edge. This temporal multigraph was then inputted into the TG-CNN model, which included a 3D CNN and an LSTM unit to process CTV3 codes and time intervals from EHRs. The 3D CNN captured

short-term temporal patterns, while the LSTM processed longer-term associations. For full technical details of the TG-CNN model and methodology, refer to Chapter 5.

6.2.1 Literature Review

When reviewing the current literature in the area the following research question was asked: What predictive methods and models are used to assess an individual's future risk of primary hip or knee replacement before secondary care referral?

Ovid MEDLINE, Scopus, Web of Science and IEEE Xplore were searched on 25/10/2024 for articles predicting hip and knee replacement risk using the following search string:

("individual risk" OR "personali$ed risk" OR "future risk" OR "replace* risk") AND ("machine learning" OR "statistical models" OR algorithms OR "predict*" OR "model" OR "risk model" OR "risk prediction" OR "AI" or "artificial intelligen*") AND ("THR" OR "THA" OR "hip arthroplast*" OR "TKR" OR "TKA" OR "knee arthroplast*" OR "hip replace*" OR "knee replace*")

Titles and abstracts were screened first, then full-texts were screened looking for papers that met the research question criteria. Papers were included if they were written in English and gave a future risk score for hip or knee replacement. Papers were excluded if risk prediction was for other outcomes rather than primary hip replacement, e.g., revision, mortality. Papers were also excluded if they explored associations or thresholding rather than developing or validating predictive models at an individual level.

Grey literature and literature reviews were included in the search to get a wide overview of current methodologies being explored. Relevant papers were extracted from reviews, using the snowballing technique.

6.2.2 Dataset

Similarly to Chapter 5, this chapter also uses NHS primary care data from ResearchOne [13]. For each of the four models produced in this chapter, the data is randomly shuffled to obtain the training (90%) and testing (10%) data sets. The data used in the hip one year in advance model of this chapter is different to the split used in Chapter 5, due to using longer patient EHRs in this chapter.

6.2.3 Patient Inclusion Criteria

Patients were required to have had at least two primary care visits within either one or five years (depending on the model) prior to undergoing a hip or knee replacement (partial or full), to ensure that graphs with at least one relationship between two visits can be established. Without two visits a EHR graph-representation cannot be established which is required for this model. Clinician-selected CTV3 codes for primary hip replacement and knee replacement were used to identify the outcome, with the incident date defined as the first occurrence of one of these codes. These codes for hip replacement are provided in Appendix B Tables B.1-B.2 and the knee replacement codes are provided in Appendix C Table C.1. The initial recorded instance of a hip or knee replacement was used as the outcome. The dataset exhibited an imbalanced case-control split, reflecting a common challenge in clinical research where, despite the high prevalence of the underlying condition, the proportion of patients undergoing hip or knee replacement is relatively small.

Patients were removed if they had a CTV3 code referring to a revision/ modification of a hip or knee replacement before a code suggesting primary replacement. If a patient had a revision for a joint but no primary joint replacement they were removed from the control group. See Figure 6.1 for the inclusion flowchart for patients included the hip one year in advance dataset.

6.2.4 Training and Test Set Formation

Two test datasets were generated for the two knee replacement models from the ResearchOne data, with a random 10% selected for the test data. 6.3% knee replacement patients were in the one-year in advance test dataset and 4.3% in the five-year in advance. For hip replacement patients, the test datasets included 7.4% replacements in the one-year in advance group and 4.5% in the five-year in advance group. The remaining data was assigned to the training set, where each joint replacement patient was matched to a control. One-to-one exact matching was performed for age, sex, and IMD to mitigate their influence on joint replacement and ensure that prediction performance was not affected by differences in these variables. The model was then trained on this balanced dataset, using 5-fold cross-validation, with one fifth of the data iteratively used as holdout/validation data to evaluate model performance. After training and optimisation, each model was recalibrated using a LR generalised linear model to adjust risk probabilities for the shift from balanced to unbalanced case-control data [228]. Complete case

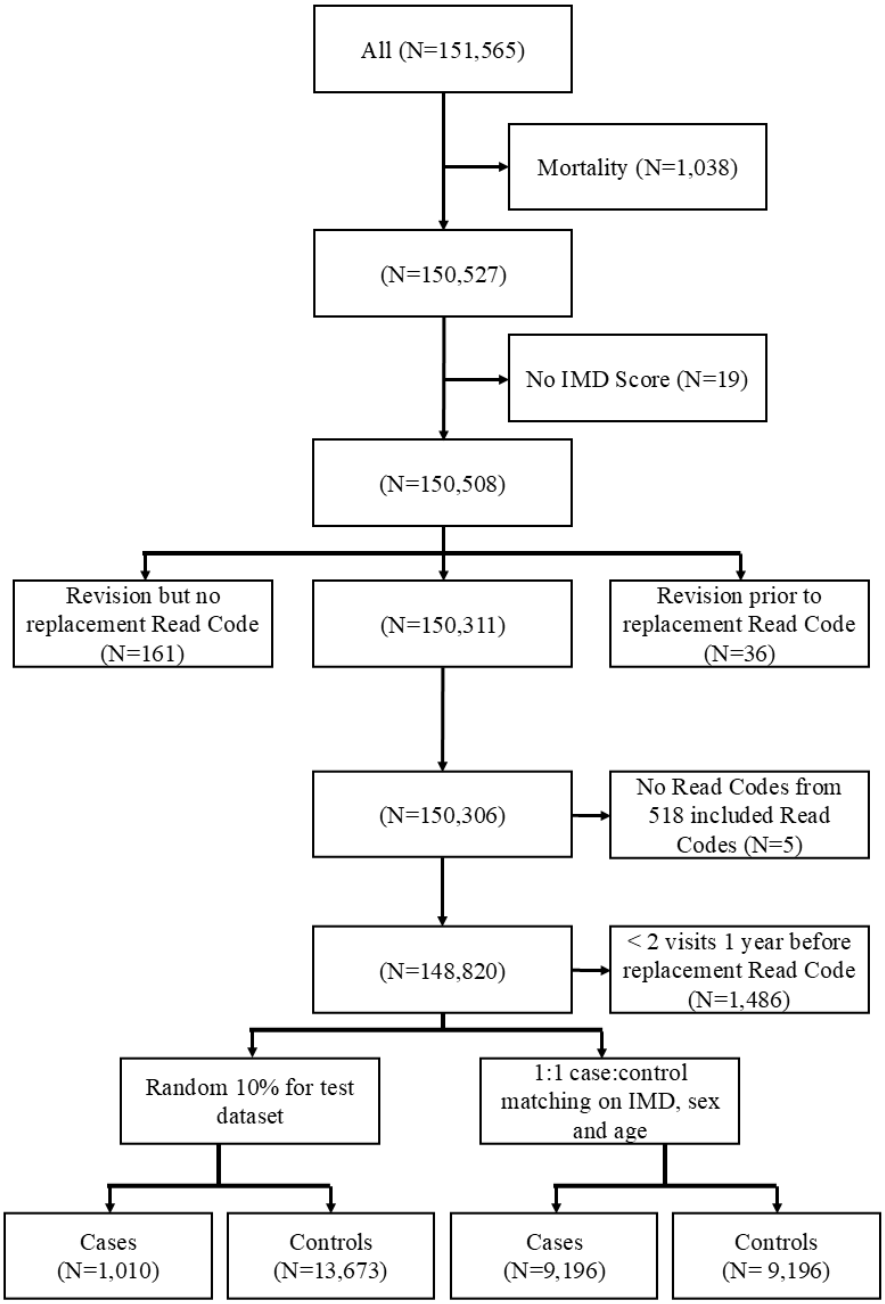


Figure 6.1: Hip one year in advance patient inclusion flowchart.

analysis was employed to estimate the marginal effect of hip or knee replacement, under the assumption of no mortality to reduce competing risks as the mortality rate was only 0.007%. If a patient dies before undergoing joint replacement, including them could introduce noise or bias as the true replacement outcome is unknown. Similarly, risk estimates could be skewed if mortality is included, particularly if death is associated with factors linked to joint replacement risk, such as serve comorbidities.

6.2.5 Model Predictors

To forecast replacements occurring one year in advance, the specific time of the replacement was identified and all available patient records from 1999 up to one year prior to the primary replacement date were gathered. Similarly, when predicting replacements five years in advance, records from 1999 to five years before the replacement event were retrieved.

The most used 512 CTV3 codes (out of 2,353 possible CTV3 codes), covering 99.46% of the recorded data, were selected to construct temporal graph-based EHR representations for each patient. Medications were appended as 6 predictors (potential graph nodes) using BNF codes, grouping drugs into opioids (04.07.02), Non-opioid Analgesics (NOAs) (04.07.01), and NSAIDs (10.01.01), these were further sub-grouped into acute and repeat prescription types. Each of the 512 CTV3 codes and 6 prescription types were mapped to a unique node $v_1, v_2, \ldots, v_{518}$, with the tensor size limited to an input of $518 \times 518 \times 200$. For demographic predictors all available demographics were used from the dataset including age, sex, and IMD. With the inclusion of 6 prescription types and 3 demographics, the total number of predictors was 521.

CTV3 hospital referral codes associated to hip or knee replacement, were excluded to prevent target leakage (where labels are included as predictors).

In Chapter 5, an ablation study was carried out to test the effect of various components on the model's performance. The best model from that study (TG-CNN w/o $\ell 2$) was used and prescription records included. Previously including prescriptions prevented the model from converging. However, including repeat and non-repeat drug prescriptions of NSAIDs, opioids and NOAs is useful, because although they could be prescribed for numerous reasons, they are a good indication of pain levels in patients.

Events occurring further in the past may have a diminishing impact on recent events, because of this models predicting patient outcomes prioritise recent events [110]. The previous chapter was

limited to using the 100 most recent primary care visits per patient to manage computational load. This chapter explores the number of time steps/visits inputted into the model, it was hypothesised that increasing the number of visits would enable the model to converge and learn more since the prescription data would not outweigh the other EHR data in the graphs. To determine the optimum number of time steps/visits the distribution of time covered by the full patient history within the analysis period was investigated, observing the difference between the number of years covered in the full history versus the last 100, 150 and 200 primary care visits. This was shown as the percentage and amount of time the EHR covered for each option. The effect on calibration from length of a patients history periods was explored to determine if temporally longer EHR histories improve model accuracy. The batch size was lowered from 128 to 64 to enable the model to run without memory errors due to the increased number of time steps included.

6.2.6 Model Evaluation

The hip and knee prediction models were tested, both one and five-years prediction in advance, with and without prescriptions, totalling 8 models. 3 baseline models: RFs (demographics, prescriptions and CTV3 codes), LR (demographics, prescriptions and CTV3 codes), and RNN models (prescriptions and CTV3 codes using cross-sectional data without elapsed time) were used as comparison models due to their limitations.

A systematic search was carried out looking for research to compare the results of the TG-CNN model to as outlined in section 6.2.1.

Calibration was used to assess risk estimate reliability, by comparing the agreement of the true number of hip/knee replacements to the predicted number of replacements. For discriminative performance the AUROC was used, which tests how well the model gives higher risks for case patients and lower risks for control patients, alongside AUPRC which quantifies the model's ability to balance precision and recall across different thresholds. The out-of-sample predictive performance (data not used in model training) was used to compare recalibrated model results.

No information regarding primary care location was provided therefore cluster analysis was not performed. However, AUPRC, AUROC and calibration curves for sex and IMD quintile subgroups for the TG-CNN models were provided.

6.3 Results

6.3.1 Literature Review Results

Inputting the search string into databases resulted in 1,613 papers being returned (Scopus (N=612), Ovid (N=369), Web of Science (N=608), IEEE Xplore (N=24)). 783 papers were removed during deduplication, leaving 830 papers for title and abstract screening. After title screening 91 papers were left. After abstract screening six papers remained. One of these papers were literature reviews [229], from which snowballing of references was performed but no extra research question relevant papers were found, totalling five full papers to screen. See Figure C.1 in Appendix C for our PRISMA flowchart [157]. In total, five papers discussed methods to give personalised future hip or knee replacement risk.

Many of the non-eligible papers were predicting risks of complications, revisions, readmissions, mortality, implant size, and other outcomes following total joint replacement, rather than predicting total joint replacement risk.

6.3.2 Study Population

More than half of the patients across all models were female, with 62.2% in the hip one year model, 61.7% in the hip five years model, 60.6% in the knee one year model, and 60.9% in the knee five years model. The mean age of patients was similar across groups, ranging from 71 to 72.5 years. The age at replacement was slightly lower, with a means between 69.2 to 70.6 years. The median age ranged from 71 to 73 years, with an age range of 43 to 90 years. The mode age was consistently between 75 and 88 years across the models.

Patients living in more deprived regions (IMD 1) visited primary care on average less frequently (54.6 $\pm$ 56.5) than people living in less deprived regions (IMD 5) (82.9$\pm$73.9). Pearson correlation analysis gave a 0.992 correlation between number of visits and IMD score, similarly there was a 0.987 correlation between number of records and IMD score. These results suggest that there is a strong correlation between IMD score and primary care utilisation. See Table 6.1 for details on average visits and records based on IMD score.

Table 6.2 displays characteristics of the cohort data included in model training and testing. Table 6.3 provides further details of characteristics grouped by training and testing dataset, with and without prescriptions.

Table 6.1: Average and median number of visits and records for each index of multiple deprivation (IMD) quintile.

IMD Quintile	Number of Records (Mean ± SD)	Number of Records (Median [min, max])	Number of Visits (Mean ± SD)	Number of Visits (Median [min, max])
1 (Most Deprived)	84.3 ± 104.1	47 [1, 1627]	54.6 ± 56.5	34 [1, 642]
2	95.0 ± 115.5	53 [1, 1710]	60.1 ± 61.6	37 [1, 880]
3	106.3 ± 125.6	60 [1, 2302]	65.6 ± 64.8	42 [1, 1005]
4	122.0 ± 135.1	73 [1, 1989]	73.4 ± 69.5	50 [1, 926]
5 (Least Deprived)	144.4 ± 150.4	92 [1, 2361]	82.8 ± 73.9	59 [1, 2217]

6.3.3 EHR Coverage

Figure 6.2 suggests that k_{max}=200 enabled the best coverage of patient EHRs with everyone having at least 25% of their record covered and less than 1% of patients having incomplete EHR coverage between 1999 and 2014. Figure 6.3 shows that on average patients had more than 6 years' worth of EHRs between 1999 and 2014.

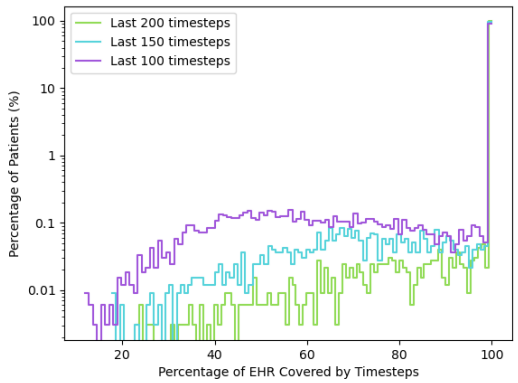


Figure 6.2: Percentage of EHR covered by including different numbers of visits (100,150 and 200) one year in advance for hip replacement prediction.

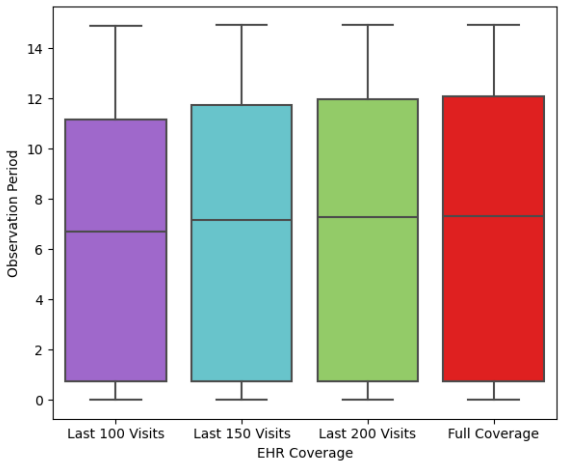


Figure 6.3: Observation period for one-year in advance hip replacement prediction, exploring coverage based on different numbers of visits. The distribution of the length of time observed within 15-year EHR data is shown when using the last 100, 150, and 200 GP visits, compared to full coverage.

In terms of full temporal coverage, when using the last 200 visits the hip one year in advance model had 98.71% coverage, the hip five year model had 98.38%, the knee one year model had 98.55%, and the knee five year model had 98.50% coverage.

The effect on calibration based on the number of years covered in the last 200 primary care

visits (one or five years prior to replacement) was explored. Figures 6.4-6.8 show the calibration curves for the four hip and knee prediction risk models. When the model is calibrated on patients with between 0-2 years worth of EHR coverage in their 200 visits the models appear to over-predict replacement risk. This effect is seen again in those with 8-15 years worth of records. The other three periods show under-prediction of the outcome risk. Patients with between 2 to 4 years worth of EHRs have the best calibration compared to the other periods. Though it is worth noting that the cohort with 8-15 years covered by their EHR mostly contains individuals with a negative outcome (no hip or knee replacement), as the last one or five years are removed from EHRs input data due to the prediction period. This makes it more likely that a patient's outcome will be over-predicted if they belong to the negative class.

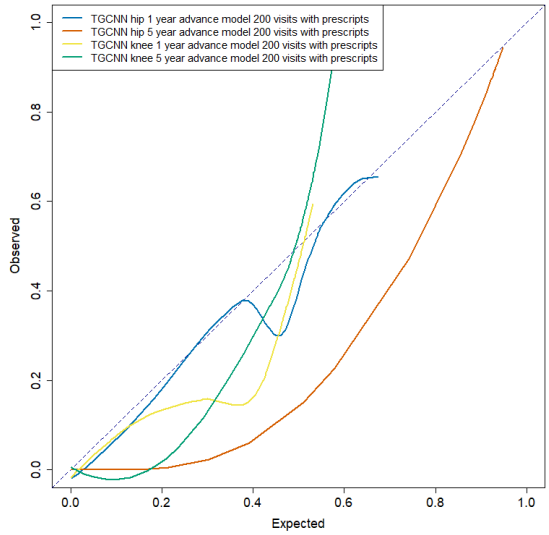


Figure 6.4: Calibration curves for individuals with 0 to 2 years of EHRs coverage.

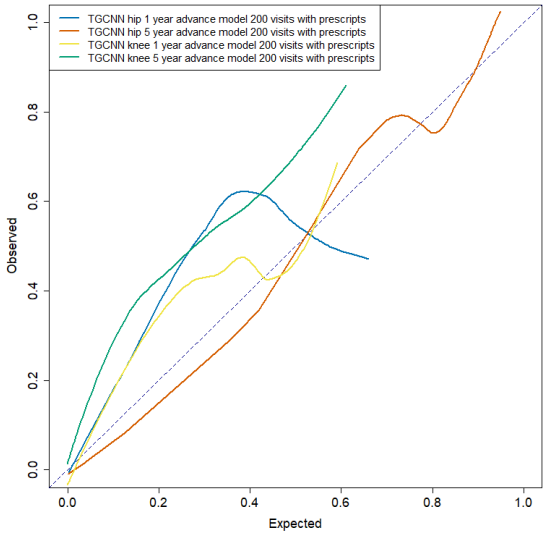


Figure 6.5: Calibration curves for individuals with 2 to 4 years of EHRs coverage.

6.3.4 Model Performances

Table 6.4 shows the AUPRC, C-slope, and AUROC values and Table 6.5 shows the PPV, sensitivity and specificity for each of the TG-CNN and baseline models. In some cases, the baseline models had better AUPRC than the TG-CNN models (LR hip one year in advance, LR knee one year in advance, Random Forest knee five years in advance, and LR knee five years in advance models). Figures 6.9-6.12 show the calibration curves from all the TG-CNN and baseline models in Table 6.4. It is vital to plot calibration curves [230], as it can be observed that the TG-CNN models fitted the population well, whilst the baseline models had poor calibration. All TG-CNN models without prescriptions, except the hip five-year advance model, predicted

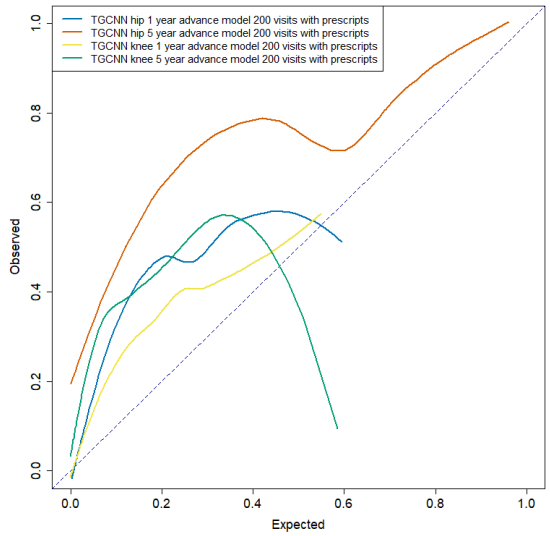


Figure 6.6: Calibration curves for individuals with 4 to 6 years of EHRs coverage.

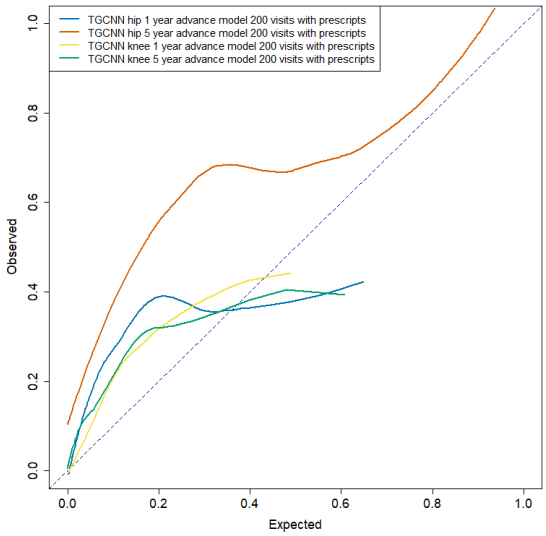


Figure 6.7: Calibration curves for individuals with 6 to 8 years of EHRs coverage.

no patients would need a replacement (PPV = 0). The hip five-year advance model had the highest sensitivity among all models. Logistic regression outperformed all models except the hip five-year advance model. The LR and Random Forest models underestimated the actual risk of needing hip or knee replacement, whilst the RNN model overestimated the actual risk. The hip five-year in advance prediction LR model had good calibration, however the other baseline models with C-slope values close to 1 were largely due to non-linear calibration. Figure 6.14 show sensitivity and specificity scores for each probability threshold for the TG-CNN and Figure 6.15 for the Logistic regression models.

All models within each prediction time and joint replacement type had non-overlapping confidence intervals when comparing AUPRC confidence intervals with and without prescriptions, suggesting there was a significant difference when including prescriptions in the model. The C-slope confidence intervals overlapped slightly in the five years in advance models for both hip and knee risk prediction.

Subgroup analysis showed that females had higher AUPRC scores than males in both the hip and knee models. For the hip one-year model, females had higher AUPRC (0.447), compared to males (0.352). The hip five-year model showed females had slightly better predicted AUPRC (0.894) versus males (0.864). In the knee one year model females had an AUPRC of 0.400, whilst males scored 0.303. The knee five-year model gave females an AUPRC of 0.481 and males 0.403. The models for the IMD 4 group often performed the best, while IMD 1 and IMD

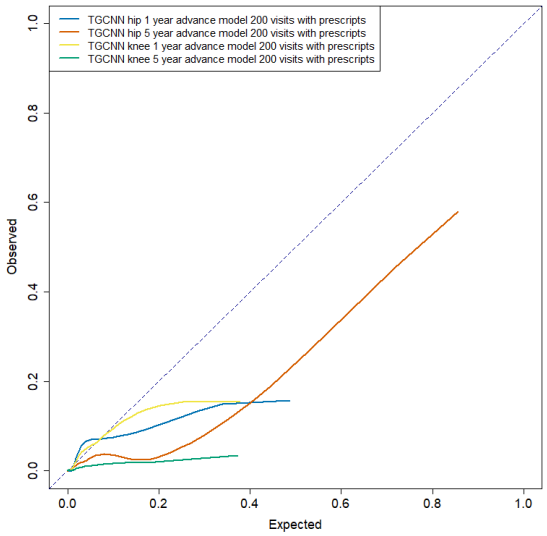


Figure 6.8: Calibration curves for individuals with 8 to 15 years of EHRs coverage.

2 frequently showed lower AUPRC scores. See Figure 6.13 for the forest plots of the subgroup AUPRCs. For subgroup calibration curves see Figures C.2-C.8 in the Appendix. In the hip one-year model, patients within the IMD 4 subgroup had the highest AUPRC (0.495), whilst IMD had the lowest (0.373). The hip five year and knee one year had the highest AUPRC in the IMD 3 subgroups (0.909 and 0.423, respectively), and lowest AUPRC in the IMD 2 subgroups (0.864 and 0.288, respectively). The knee five-year model had the highest AUPRC in the IMD 4 subgroup (0.530), and the lowest AUPRC scores in the IMD 2 subgroup (0.346). The hip five-year model had the best generalisability across subgroups, with all subgroups having AUPRC scores over 0.864. The knee one year model had the next smallest AUPRC subgroup range between 0.288 and 0.430, followed closely by the hip one-year model (0.352-0.495), and then finally the knee five-year model (0.346-0.530).

Five papers were identified as relevant to our systematic search research question [15, 231, 232, 233, 234]. The model with the best performance for hip and knee replacement risk was the Oxford Knee and Hip Score (OKHS) model [231], with AUROCs of 0.87 and 0.83 for hip and knee prediction. The TG-CNN model outperformed the current state of the art risk prediction tools for hip and knee replacement by up to 0.1 and 0.13 (AUROCs), respectively.

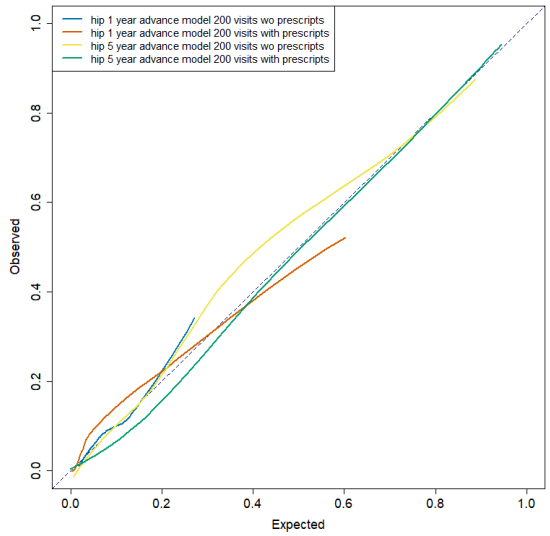


Figure 6.9: Calibration curves for the TG-CNN models for hip replacement risk.

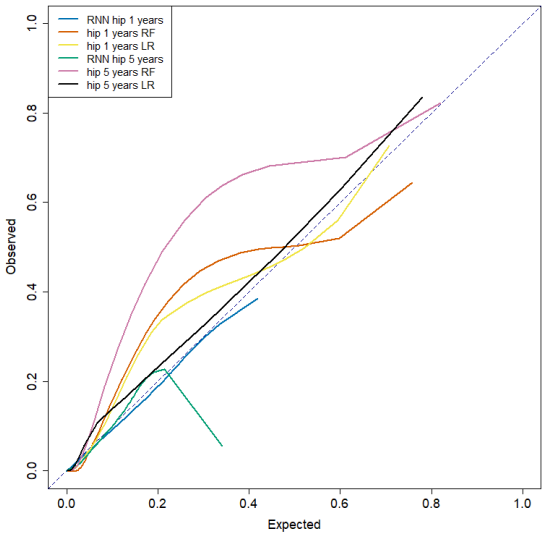


Figure 6.10: Calibration curves for the baseline models for hip replacement risk.

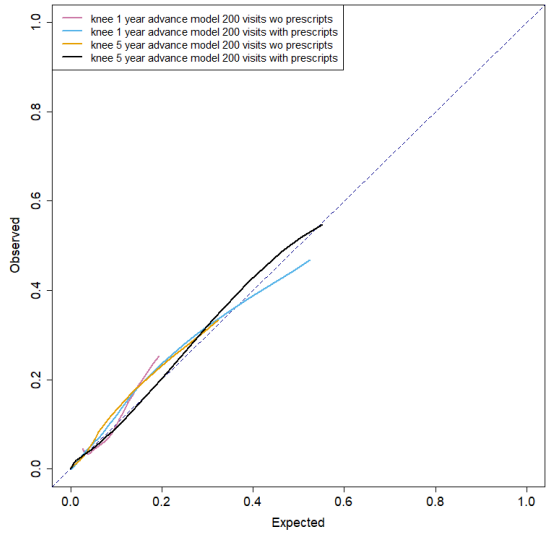


Figure 6.11: Calibration curves for the TG-CNN models for knee replacement risk.

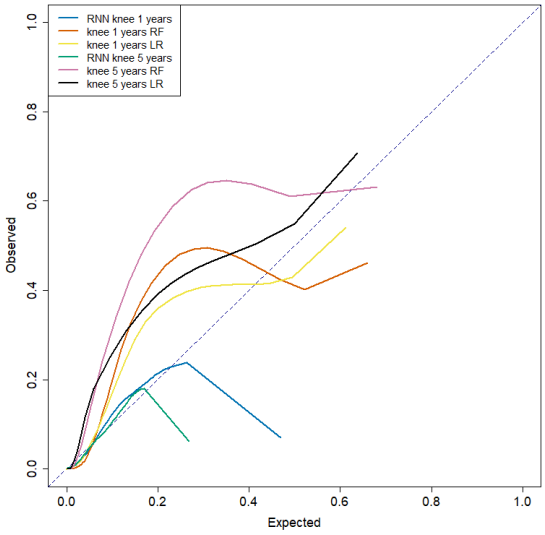


Figure 6.12: Calibration curves for the baseline models for knee replacement risk.

Table 6.2: Replacement prediction model cohort characteristics. The BMI statistics are derived from the last recorded BMI measurement of each patient. The first recorded IMD value for each patient was used, given its stability over time. BMI=body mass index, IMD=index of multiple deprivation, std=standard deviation.

Characteristic	**Hip 1 Year (N=33,074)**	**Hip 5 Years (N=24,732)**	**Knee 1 Year (N=30,492)**	**Knee 5 Years (N=24,370)**
Sex: female (N (%))	20,562 (62.2)	15,252 (61.7)	18,482 (60.6)	14,837 (60.9)
Hip/Knee replacements (N (%))	10,206 (30.9)	5,857 (23.7)	8,723 (28.6)	5,543 (22.7)
Age, y				
mean (std)	72.5 (9.9)	71.4 (10.2)	72 (9.8)	71 (9.9)
median (min,max)	73 (53,89)	71 (53,89)	72 (53,89)	71 (53,89)
mode	88	88	88	66
Age at replace, y				
mean(std)	69.2 (8.5)	70.6 (8.5)	68.9 (8.1)	69.6 (8.2)
median (min,max)	70 (43,90)	71 (48,90)	69 (46,89)	70 (47,89)
mode	77	77	75	75
BMI (N (%))				
Severely obese	1,017 (3.1)	792 (3.2)	1,205 (4.0)	943 (3.9)
Obese	9,480 (28.7)	7062 (28.6)	9,538 (31.3)	7,474 (30.7)
Overweight	12,208 (36.9)	9101 (36.8)	10,948 (35.9)	8,775 (36.0)
Healthy weight	8,193 (24.8)	6052 (24.5)	6,855 (22.5)	5,578 (22.9)
Underweight	544 (1.6)	416 (1.7)	437 (1.4)	367 (1.5)
Missing	1,632 (4.9)	1309 (5.3)	1,509 (4.9)	1,233 (5.1)
IMD, quintile (replacements / total (%))				
1 (most deprived)	2749/8301 (33.1)	1588/6104 (26.0)	2021/7007 (28.8)	1280/5611 (22.8)
2	2546/7836 (32.5)	1452/5753 (25.2)	2035/6966 (29.2)	1282/5515 (23.2)
3	2246/7125 (31.5)	1267/5128 (24.7)	1937/6544 (29.6)	1234/5172 (23.9)
4	1493/5116 (29.2)	852/3876 (22.0)	1435/5002 (28.7)	919/4057 (22.7)
5 (least deprived)	1172/4696 (25.0)	698/3871 (18.0)	1295/4973 (26.0)	828/4015 (20.6)
Events Recorded				
mean(std)	58.0 (91.8)	61.9 (97.5)	66.5 (99.8)	63.9 (95.4)
median (min,max)	23.0 (2,1890)	24.0 (2,1255)	29.0 (2,1494)	29.0 (2,1890)
mode	2	2	2	2

Table 6.3: Extra dataset information. Where 'max # records' is the maximum number of records a single patient has. CV=cross-validation data set, SD=standard deviation, w drugs=with drugs, w/o=without, IMD=index of multiple deprivation score.

	Knee 1 year	**Knee 5 years**	**Hip 1 year**	**Hip 5 years**
# with a revision but no replacement	71	71	161	161
# with a revision before a primary replacement code	31	31	36	36
# patients with mortality	1,036	1,036	1,038	1,038
# patients with no IMD score	19	19	19	19
# patients excluded due to code coverage	5 / 150,404	5 / 150,404	5 / 150,305	5 / 150,305
# replacement patients before windowing	10,243	10,243	12,706	12,706
# replacement patients after windowing	9,323	6,806	11,220	7,551
# patients in CV set (case | control)	7,851 | 7,851	4,949 | 4,949	9,196 | 9,196	5,243 | 5,243
# patients in test set (before halving) (case | control)	872 | 6,959	594 | 6,939	1,010 | 13,673	615 | 13,633
Full time coverage	98.55%	98.50%	98.71%	98.38%
CV set max # records (w/o drugs)	210	154	218	154
CV set mean (SD) # records (w/o drugs)	7 (10)	5 (7)	6 (9)	5 (6)
CV set median records (w/o drugs)	3	2	3	2
CV set max # records (w drugs)	524	303	471	246
CV set mean (SD) # records (w drugs)	23 (36)	14 (23)	17 (31)	11 (20)
CV set median records (w drugs)	5	3	4	3
Test set max # records (w/o drugs)	191	399	233	399
Test set mean (SD) # records (w/o drugs)	18 (15)	18 (16)	18 (15)	18 (16)
Test set median records (w/o drugs)	14	14	13	14
Test set max # records (w drugs)	622	741	889	620
Test set mean (SD) # records (w drugs)	47 (54)	47 (54)	47 (56)	49 (55)
Test set median records (w drugs)	26	27	26	27

Table 6.4: AUPRC, C-slope and AUROC results for the models on the unseen test data set. AUPRC thresholds based on prevalence for each dataset are as follows: hip one year = 0.069, hip five years = 0.227, knee one year = 0.059, knee five years = 0.041 (AUPRC scores lower than their respective threshold can be deemed as uninformative models). The best scores for each replacement and year in advance type for each performance metric (columns) are given in bold. Prescriptions (prescript). RF=Random Forest.

Model	**AUPRC [95% CI]**	**C-slope [95% CI]**	**AUROC [95% CI]**
Hip 1 year in advance			
TGCNN with prescript	0.409 [0.366, 0.451]	0.976 [0.970, 0.983]	0.919 [0.918, 0.920]
TGCNN w/o prescript	0.192 [0.164, 0.219]	1.075 [1.065, 1.084]	0.734 [0.732, 0.736]
RNN with prescript	0.293 [0.260, 0.325]	1.050 [1.040, 1.061]	0.895 [0.894, 0.896]
RF with prescript	0.431 [0.389, 0.472]	**0.996 [0.990, 1.001]**	0.933 [0.932, 0.933]
LR with prescript	**0.477 [0.428, 0.525]**	0.995 [0.989, 1.001]	**0.938 [0.937, 0.938]**
Hip 5 years in advance			
TGCNN with prescript	**0.879 [0.833, 0.924]**	1.047 [1.034, 1.061]	0.967 [0.966, 0.968]
TGCNN w/o prescript	0.762 [0.704, 0.820]	1.059 [1.046, 1.072]	0.913 [0.911, 0.915]
RNN with prescript	0.184 [0.158, 0.209]	1.073 [1.059, 1.088]	0.894 [0.893, 0.895]
RF with prescript	0.567 [0.507, 0.627]	0.969 [0.961, 0.978]	0.969 [0.968, 0.969]
LR with prescript	0.620 [0.564, 0.676]	**1.021 [1.010, 1.032]**	**0.973 [0.973, 0.973]**
Knee 1 year in advance			
TGCNN with prescript	0.353 [0.302, 0.403]	**0.997 [0.989, 1.005]**	0.915 [0.914, 0.916]
TGCNN w/o prescript	0.140 [0.115, 0.165]	1.094 [1.077, 1.110]	0.661 [0.658, 0.664]
RNN with prescript	0.193 [0.171, 0.216]	0.901 [0.892, 0.910]	0.864 [0.863, 0.866]
RF with prescript	0.335 [0.294, 0.376]	0.953 [0.948, 0.958]	0.925 [0.924, 0.926]
LR with prescript	**0.374 [0.326, 0.423]**	0.965 [0.959, 0.972]	**0.928 [0.927, 0.929]**
Knee 5 years in advance			
TGCNN with prescript	0.442 [0.382, 0.503]	0.980 [0.970, 0.991]	0.955 [0.954, 0.956]
TGCNN w/o prescript	0.238 [0.199, 0.277]	0.990 [0.981, 0.999]	0.856 [0.853, 0.858]
RNN with prescript	0.146 [0.126, 0.166]	0.951 [0.938, 0.965]	0.860 [0.859, 0.862]
RF with prescript	0.467 [0.410, 0.524]	**0.997 [0.989, 1.004]**	0.963 [0.963, 0.964]
LR with prescript	**0.500 [0.438, 0.561]**	0.975 [0.966, 0.984]	**0.965 [0.965, 0.966]**

Table 6.5: PPV, sensitivity and specificity results for the models on the unseen test data set. RF=Random Forest, LR=logistic regression.

Model	PPV	Sensitivity	Specificity
Hip 1 year in advance			
TGCNN with prescript	0.474	0.232	0.981
TGCNN w/o prescript	0.000	0.000	1.000
RNN with prescript	0.100	0.002	0.999
RF with prescript	0.497	0.194	0.986
LR with prescript	0.586	0.255	0.987
Hip 5 years in advance			
TGCNN with prescript	0.822	0.836	0.947
TGCNN w/o prescript	0.752	0.655	0.937
RNN with prescript	0.000	0.000	0.999
RF with prescript	0.726	0.319	0.995
LR with prescript	0.725	0.336	0.994
Knee 1 year in advance			
TGCNN with prescript	0.528	0.108	0.994
TGCNN w/o prescript	0.000	0.000	1.000
RNN with prescript	0.000	0.000	0.996
RF with prescript	0.392	0.128	0.987
LR with prescript	0.550	0.126	0.994
Knee 5 years in advance			
TGCNN with prescript	0.525	0.209	0.992
TGCNN w/o prescript	0.000	0.000	1.000
RNN with prescript	0.000	0.000	1.000
RF with prescript	0.649	0.205	0.995
LR with prescript	0.680	0.236	0.995

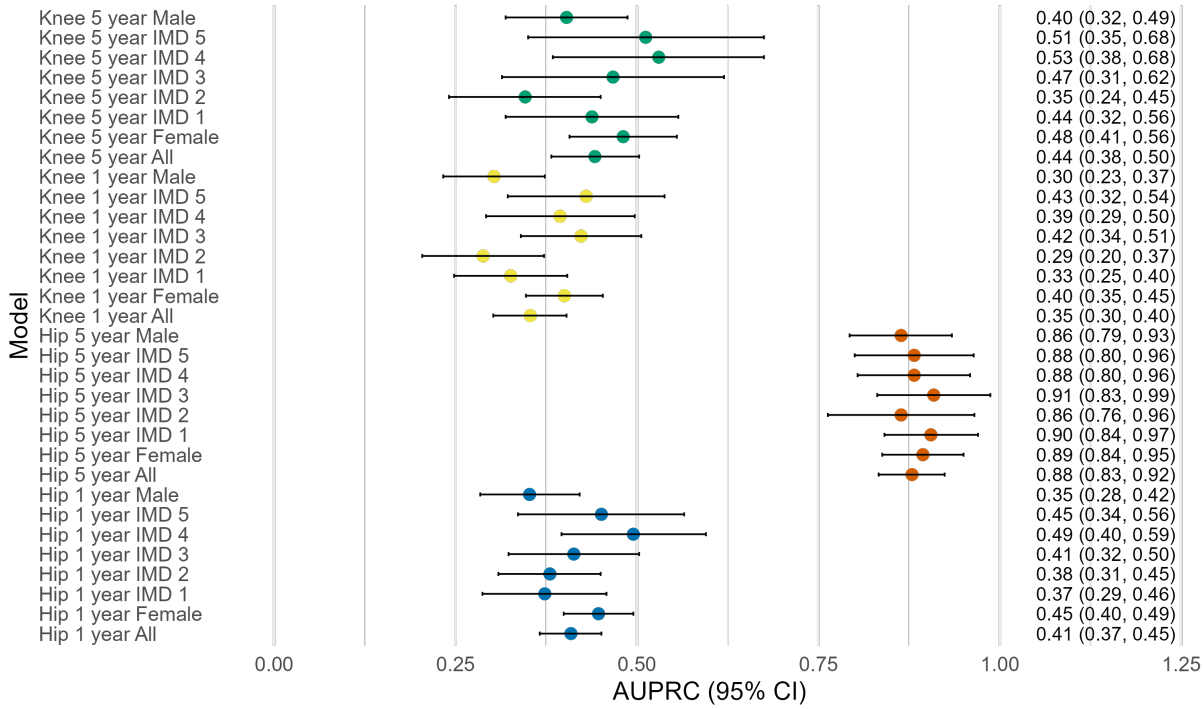


Figure 6.13: Forest plot showing area under the precision recall curve (AUPRC) scores and 95% confidence intervals for each of the TG-CNN models (with prescription data included) and subgroups.

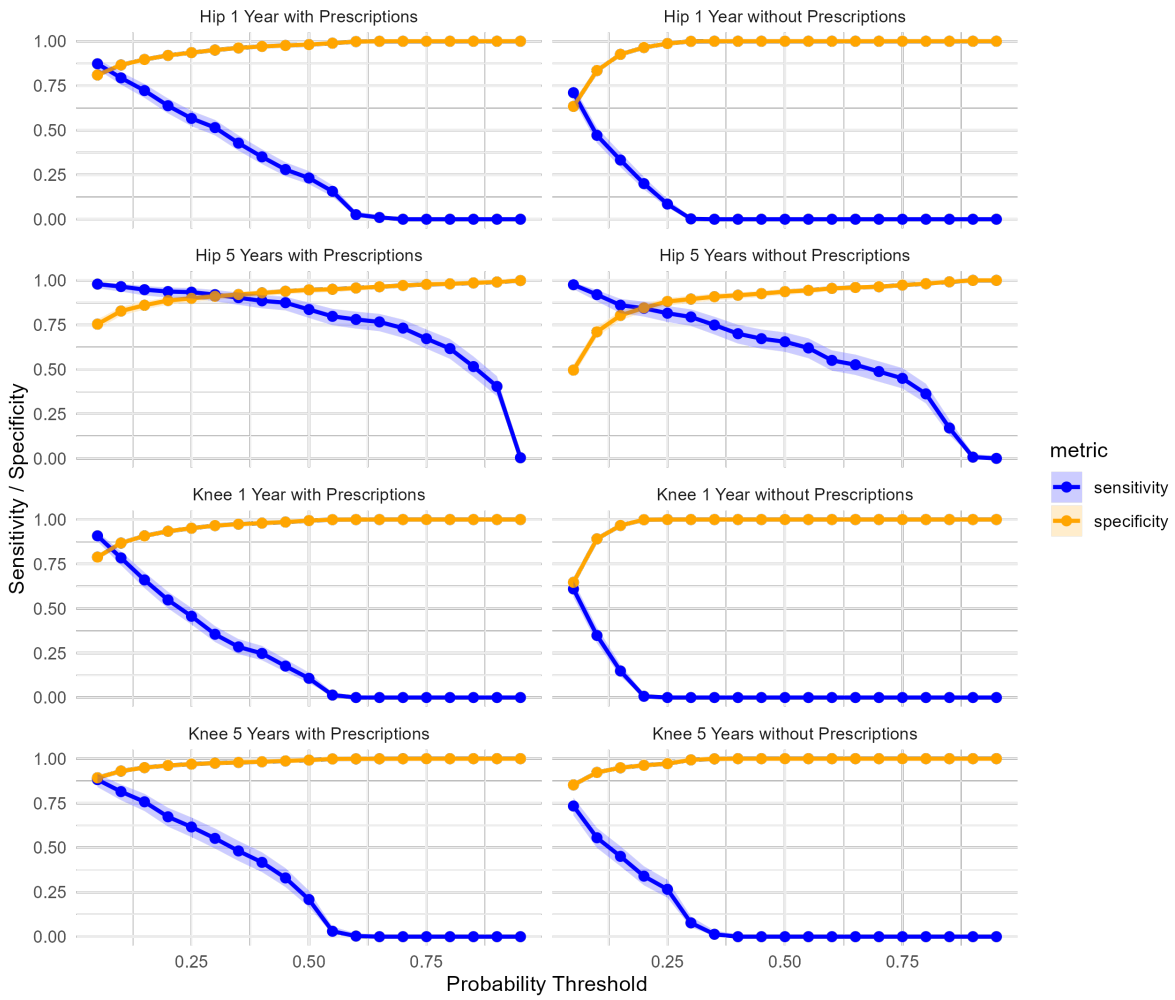


Figure 6.14: Sensitivity and specificity plots for the TG-CNN models at each probability threshold. The hip five years in advance models have the best sensitivity and specificity across the thresholds.

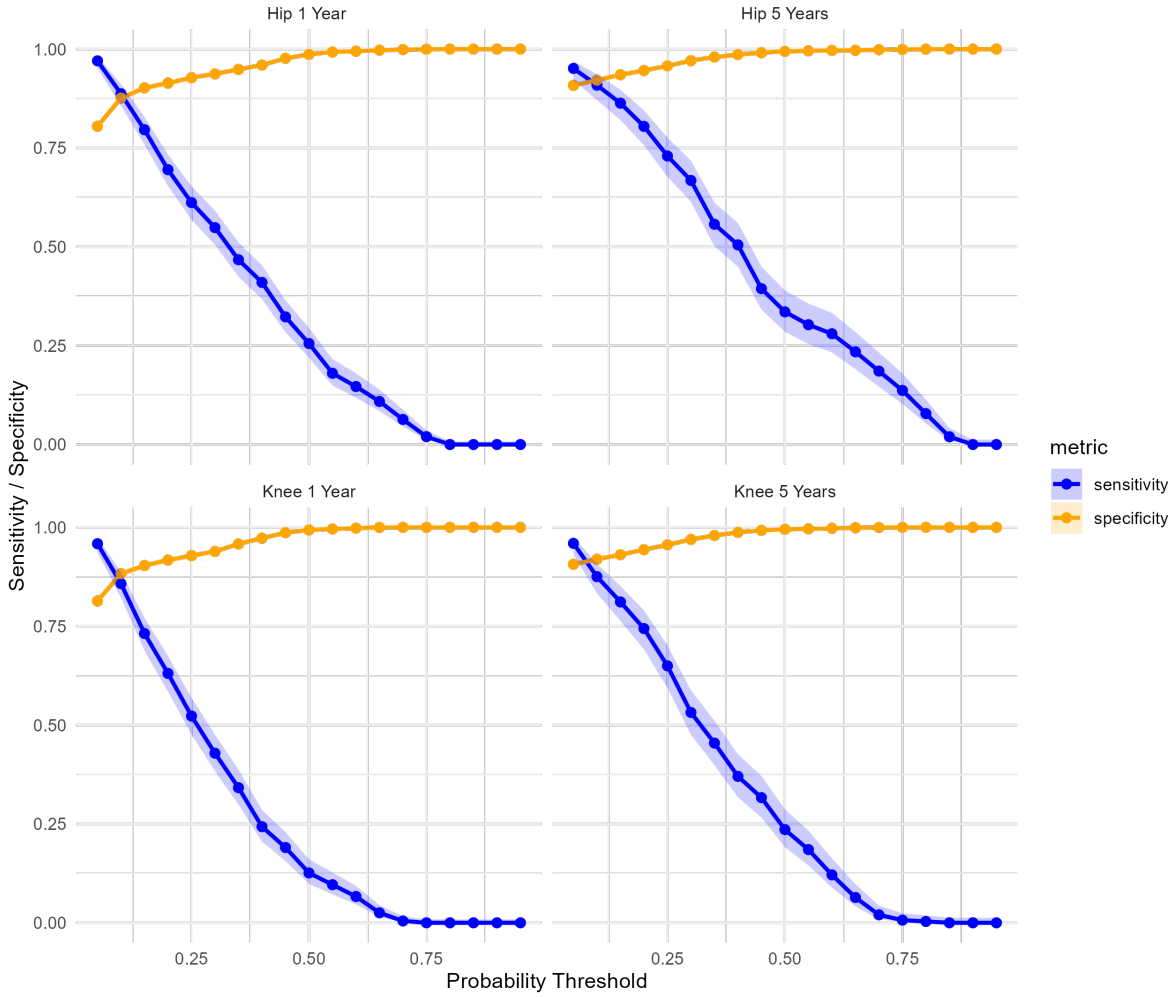


Figure 6.15: Sensitivity and specificity plots for the Logistic regression models at each probability threshold.

6.4 Discussion

This study demonstrates that hip and knee replacement risk can be predicted one or five years in advance using CTV3 codes, demographic, and prescription data from individual patient history using up to 200 of their historical primary care visits. Predicting replacement risk five years in advance gave better predictive performance than one-year in advance in all models except the RNN model. Prediction five years in advance in the hip model may have improved performance due to higher hip replacement prevalence in this group (0.069) compared to the one-year group (0.227). The TG-CNN produced the best calibration curves for hip and knee replacement risk prediction.

The performance of the models predicting joint replacement five years in advance was expected to perform worse than the one year in advance model, due to having less data and patterns to pick up relevant to replacement risk prediction. This could be due to the five years in advance models having less noise, more long-term predictors, reduced overfitting, or it being able to focus on long-term health deterioration which may be a clearer target for joint replacement prediction.

Subgroup analysis revealed that the model performed slightly worse on males than females, this could be in part due to there being a higher population of women who have joint replacements. Females could also have clearer trajectories aligning to future joint replacement requirement due to hormonal and bone mass density changes [235], leading to higher rates of osteoporosis [236]. Previous association studies have also found that being female and being over 60 increases risk of knee replacement [50]. Subgroup analysis also showed that patients in the IMD 4 group often had the better performing model, whilst the patients living in the more deprived regions (IMD 1 and 2) had worse performance. Hip and knee replacement prevalence was mostly higher in the IMD 1 and 2 groups, suggesting that it was harder to identify individuals at risk of requiring a joint replacement in these groups. The data used in this study showed more frequent visits and larger numbers of records in less deprived population groups. This could be because people living in more deprived areas may have less access to primary care facilities, leading to less EHR data for these patients [237, 238]. Future work of outcome pathways relative to subgroups could be carried out using causal inference alongside predictive models. Correcting for class imbalance by resampling has been shown to over-estimate risk and may not improve estimations during re-calibration on some models [239], however the TG-CNN model showed

promising performance improvement after re-calibration on the out-of-sample imbalanced data despite training on balanced data. Additionally, though re-calibration can improve calibration, discrimination is not effected [239].

6.4.1 Related Work

Over 20 models have been developed to predict hip (N=4), knee (N=22), or generic joint (N=1) replacement using secondary care data. Predictors in these models include: radiographic details of structural tissue damage (e.g., cartilage, bone marrow and meniscus) [27, 63, 220, 234, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256], ultrasonography [27], demographic data [27, 63, 234, 240, 241, 244, 251, 253, 257], NSAID use [63], occupation [63], physical activity [63, 244], knee alignment [63], knee extensor and flexor strength [63], Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) score [63, 234, 252, 257, 258, 259], OA (Kellgren-Lawrence grade $\geq$ 3) [63, 220, 234, 241, 244, 246, 247, 258], Comorbidities [63], anterior cruciate ligament rupture [63], clinical symptom data [27, 234, 241, 244, 247, 248, 251, 254, 257, 258, 260], hip shape [261], and stability [254]. Models used to predict hip and knee replacement using secondary care data included: Cox Proportional Hazards [27, 234, 240, 243, 244, 255, 256], LASSO [253, 262], RFs [262], CNNs [252, 262], RNNs [262], LR [63, 220, 241, 242, 246, 248, 250, 254, 261], deep CNNs [249, 260], clustering [257], unsupervised data augmentation [245], gradient boosting machine [251], DeepSurv [247], and artificial neural network [258] models. These models had predictive AUROC performances between 0.79-0.94, with the hip replacement risk 9-years in advance deep CNN model [260] having the best performance.

The Oxford Knee and Hip Score (OKHS) was used to screen patients for surgical or non-surgical hip and knee referrals through brief patient surveys regarding pain and mobility, achieving AUROC scores of 0.83 and 0.87 for knee and hip referral types, respectively [231].

A Fine and Gray competing risk prediction model was developed using factors such as age, previous knee replacement, non-replacement knee surgery, OA prescriptions, comorbidity index, and mental health to estimate the five-year risk of total knee replacement, yielding an AUROC of 0.67 [232].

Cumulative incidence equations were applied to estimate the 10-year risk of requiring a hip or knee replacement in individuals with OA and/or hip or knee pain. The model was refined to 20

predictors for hip replacement and 24 for knee replacement, resulting in an AUROC of 0.72 for hip replacement and 0.78 for knee replacement. Strong predictors of increased knee replacement risk included oral NSAIDs, opioid analgesic prescriptions, intra-articular injections, and prior arthroscopic knee surgery within three years before the index consultation, while a history of hip injury strongly indicated future hip replacement risk [233].

For patients with OA, a knee replacement risk model using Random Survival Forests achieved an AUROC of 0.807 by incorporating predictors such as sex, age, education level, BMI, occupational activity, smoking, history of stroke, diabetes, heart attack, pain medication, knee injury and surgery, knee pain and stiffness, and walking aid use. SHAP identified pain medication, age, surgery, diabetes, and BMI as the highest risk factors for knee replacement [263].

Total knee replacement risk was also modelled using separate primary care and secondary care predictors in patients with knee pain in the previous three months. Cox proportional hazard models, based on demographic data (age, sex, race), knee arthroscopy history, knee pain, analgesics use, glucosamine use, BMI, and WOMAC pain score for the primary care model, and on demographics, knee pain, analgesics use, WOMAC pain score, and Kellgren-Lawrence grade for the secondary care model, showed that the secondary care model significantly outperformed the primary care model, with an AUROC of 0.87 compared to 0.79 [234].

The TG-CNN models improved upon the existing hip and knee replacement risk prediction models in this area by considering additional predictors such as IMD score, prescriptions, and 512 CTV3 codes. Also, the focus was extended beyond only OA patients or patients with recent joint pain. Whilst OA and joint pain hugely contribute to joint replacement, they are not the only valuable predictors. Increasing the number of predictors (in our cases CTV3 codes) helps cover a wider range of patient trajectories and making it less likely to miss patients that may have more unusual pathways to replacement. Primary hip and knee replacement for both partial and full replacement was covered in this work, rather than just total replacement. Using graphs enables incorporation of temporality in pathways, such that CTV3 codes can be repeated at different time steps rather than a binary indicator each predictor. For example, by having a painkiller prescription recorded across multiple visits pain severity can be estimated, as pain is associated to hip replacement procedures [264]. The TG-CNN model provides individual trajectory importance, i.e. what parts of a patient's pathway has led to the predicted risk outcome, rather than a global predictor importance, to provide clearer explainability and help

inform clinical decision making. Predictors covering mental health conditions such as anxiety and depression were also included.

6.4.2 Limitations

One of the largest limitations of this work is the lack of external validation, this model has not been tested on an external dataset, however using out-of-sample data on the re-calibrated model gives suggests good performance on new data.

The dataset used did not contain ethnicity information which is a beneficial predictor of hip replacement as some ethnic groups are more prone to specific conditions. This could be informative to give clinicians and patients explainability of the model's risk decision. Only CTV3 codes and limited demographic information were used, however including images and/or imaging reports may be more informative to replacement prediction [234, 261], though the benefit of using CTV3 codes is minimisation of invasive or costly screening methods by predicting replacement at primary care level earlier and before severe degradation occurs. The regions or practices the ResearchOne dataset cover were not provided, and this dataset is smaller than datasets such as CPRD. These models were also trained on true replacement dates, rather than optimum surgery times, this meant that waiting times were not considered.

CTV3 codes are no longer used, however this model could be retrained or revalidated in future research with alternative clinical codes using the same methodology as provided in this thesis. While the 512 CTV3 codes covered 99.46% of recorded clinical events, we did not assess the impact of using fewer codes. However, given the high coverage, rerunning the analysis was unlikely to provide additional insights.

Obesity significantly impacts OA progression, co-morbidities, and surgical decisions due to interconnected physiological and behavioural factors. Excess weight increases mechanical stress on weight-bearing joints, particularly the knees and hips, accelerating cartilage degradation and contributing to OA development [265]. Obesity also affects blood sugar regulation, with insulin resistance and chronic inflammation impairing cartilage repair mechanisms, further exacerbating joint deterioration [266]. Obesity is also linked to cardiovascular fitness, which influences surgical eligibility and recovery outcomes. Patients with obesity often experience reduced mobility, limiting their ability to engage in joint-preserving physical activity. However, certain high-impact exercises may also accelerate joint damage, therefore patients should partake in activities

suitable for their bodies to prevent harm [265]. Obesity contributes to a circular relationship between mood, pain, sleep, and physical activity. Chronic pain can lead to poor sleep and depression, which in turn reduces motivation for physical activity, worsening OA symptoms [267]. These complex, often rhythmic exposures are largely unseen in traditional training data, making models for hip and knee replacement prediction challenging.

6.4.3 Future Work

In future work, this model could be externally validated to assess generalisability. Next, conducting a meta-analysis of various interventions could help assess their costs and associated risks, potentially identifying more cost-effective approaches. Additionally, involving a Patient and Public Involvement (PPI) group for risk grouping, alongside performing decision curve analysis, could further demonstrate whether these strategies are viable for reducing healthcare costs. The TG-CNN model could then be utilised for triaging patients for follow-up care. The model could be integrated into primary care to return a risk score for the patient of hip or knee replacement if a joint pain CTV3 code is inputted. The next chapter delves into explainable methods to visually show clinicians which parts of the EHR are associated to the risk score produced by the TG-CNN model.

Early prediction increases clinical utility for treatment and care planning. However, the one year in advance model may also be useful for short term risk, for example in patients experiencing severe pain or reduced mobility, their clinician may wish to observe the likelihood of them needing a replacement sooner vs later.

While subgroup analyses were conducted across age, sex and IMD, other biases may be present in the training data that could affect the fairness of the model across populations. These biases can arise from data collection methods, representation of different groups, and model characteristics. The distribution of age in the training data may not reflect the true population. If age groups are over or underrepresented the model may perform differently across age subgroups, leading to biased predictions. If sex imbalance exists, the model may be biased and favour the overrepresented sex. If certain deprivation levels are underrepresented, or the diversity of socioeconomic status is not captured sufficiently, the model may struggle to predict fairly across IMD groups. Further work is required to monitor and address model prediction biases, particularly in new out-of-sample data.

Causal inference could enhance the subgroup analysis carried out in this chapter. Applying causal inference methods allows us to consider whether factors influence the likelihood of joint replacement, rather than just looking at the subgraphs that the model associates to the prediction. These methods help find modifiable risk factors, enabling healthcare professionals to assess whether interventions, like weight loss programs, could meaningfully reduce the probability of surgery. By mitigating confounding biases through techniques like propensity score matching or instrumental variable analysis, causal inference ensures that subgroup analysis yields more reliable and actionable insights. Causal inference can also guide personalised treatment strategies and inform healthcare policies. By analysing subgroups, we can determine how different patient populations respond to specific interventions. This enables clinicians to tailor treatment approaches to maximize effectiveness for each demographic.

6.5 Conclusion

In conclusion, this work demonstrates that hip and knee replacement risks can be effectively predicted up to five years in advance using patient-specific data, including CTV3 codes, demographic information, and prescription history. The use of the TG-CNN model proved impactful, outperforming existing risk prediction tools by 5% and 9% for hip and knee replacement predictions, respectively, while also producing the best calibration curves. Notably, predictions made five years in advance generally exhibited superior performance compared to one-year predictions, except when using the RNN model.

Advancements of this work include the incorporation of a broader range of predictors, such as the IMD score and mental health conditions, extending the scope beyond traditional predictors like osteoarthritis and joint pain. This ensures a more inclusive assessment of patient trajectories, reducing the likelihood of overlooking less conventional pathways to joint replacement. The inclusion of temporal data through graph-based modelling further enhanced predictive accuracy by capturing the progression of conditions such as pain over time. By covering both partial and total joint replacements, this study broadens its clinical relevance and demonstrates the utility of advanced predictive models in enhancing hip and knee replacement risk assessment and patient management.

Chapter 7

Explainable Methods for the TG-CNN Model

7.1 Introduction

Previous chapters sought to predict hip and knee replacement risk in advance using clinical codes from EHRs in TG-CNN, with good performance. However, these methods currently lack explainability and do not provide clinicians with insights into why the model generated a particular risk score, which makes models less trustworthy for clinical decision-making.

Explainability in the context of ML can be defined as the inner working of a model being explained in understandable terms to highlight the factors which led to the models decision. Many AI models are "black-box" meaning that often engineers and users do not know what is going on internally, making them untrustworthy due to non-transparency. An algorithm may be used to assist a healthcare decision, but AI will never be allowed to run completely autonomously for decision making in healthcare, explainability should be provided to allow the clinician and (where appropriate) the patient to understand model conclusions [268]. This is vital to ensure patient safety and fair decision making, whilst empowering patients to give them an understanding of their health and the decisions they face. The GDPR states that individuals can demand explanations about the logic, and consequences of automated decisions under certain circumstances (e.g., profiling, Article 22). Articles 13, 14, and 15 indirectly support transparency and an individual's ability to understand and challenge decisions made by automated tools [11]. The main goal of explainable AI models in healthcare should be to

explain prediction decisions and behaviours to non-experts in ML, to be trusted in clinical settings.

Making the TG-CNN model explainable provides a solution to the four major challenges outlined by Xie et al., 2022 [269]: The TG-CNN model accounts for variable time intervals between primary care visits, it uses sparse linear algebra and graph representations to mitigate for EHR data sparsity, individual patient graphs are created to deal with data heterogeneity, and the methods described in this chapter enable model opacity.

The contribution of this chapter are as follows:

1. Four post-hoc explainability methods are used to visually show TG-CNN prediction decision influence for individual patients.
2. Interactive graph visualisations are created which allow clinical users to interact with historical patient EHRs.
3. These interactive graphs are evaluated using human/clinician evalulation feedback alongside Edge Detection Bias (EDB), sensitivity, and sparsity.
4. Subgraph analysis is performed to find frequent subgraphs which impact model decision.

7.2 Related Work

Zhao et al., 2023 provided an overview of multiple explainability methods used for deep time-series models [270]. Post-hoc explainability methods include backpropagation-based (e.g. Grad-CAM) or importance based to see things such as layer impact on classification accuracy [9], approximation-based, (e.g. LIME and SHAP), and perturbation-based methods (e.g. erasure, ablation and permutation) [270, 271, 272]. Ante-hoc methods include attention-based, causality-based, and physics rule-based models. Kakkad et al., 2023 summarises the explainability for GNNs, splitting methods into factual (explanations using input feature influence on prediction) and counterfactual (providing explanation by finding how much change is needed in the input graph to change the prediction) methods [273]. Liu et al., 2022 categorise methods for explaining GNNs: Model-agnostic methods including subgraph-based approaches, which identify influential subgraphs [274], feature attribution methods that assess node/ edge importance and their interactions [272], using techniques like LIME and SHAP, and counterfactual explanations

that explore minimal changes leading to different predictions. Model-specific methods include attention-based models, graph masking approaches, and self-explaining GNNs that integrate explanation generation into their predictions. Attention mechanisms are popular to improve neural network interpretability. They compute a weighted context vector to represent the input sequence, with weights indicating the importance of different segments. Attention mechanisms are often integrated into a model's structure. Post-hoc explanation methods involve gradient-based approaches for interpreting feature importance, decomposition methods like Layer-wise Relevance Propagation for attributing output scores, and surrogate models that approximate GNN behaviour using simpler, interpretable models. Instance-level explanations focus on node, edge, or graph-level explanations.

In terms of work similar to this chapter, Lauritsen et al., 2020 used temporal convolution networks using sequences of EHR information to find features relevant to sepsis, acute kidney injury or acute lung injury [9]. For their explainable AI method they used layer-wise relevance backpropagation to highlight relevance. Visually they describe this using circles for each EHR finding over time, with value intensity such as Kidney eGFR being indicated by a colormap, and parameter relevance being shown using circle size. Ying et al., 2019 created the GNNExplainer method to provide explainability of GNNs [274]. GNNExplainer works to find the most important subgraph and node features that contribute to the model prediction, it works by finding the smallest subgraph possible whilst keeping the same prediction decision. Node and edge contribution are quantified and shown using heatmaps.

Contrastive gradient-based saliency maps generate heatmaps by differentiating the output of the model with respect to the model input, where the positive values in the heatmap represent the input feature importance [275]. Class Activation Mapping (CAM) provides a slight modification for CNNs by taking the gradients of the output of the model with respect to the output of the final CNN layer; this is thought to return more meaningful features in the input data. However, CAM requires the final CNN layer to be just before the final output layer of the model, with only a global average pooling layer in-between. Grad-CAM improves on this further by allowing layers between the final CNN layer and the output by weighting feature maps, whilst also considering feature map activations rather than just gradients [275, 276].

Pope et al. 2019 adapted explainability methods for CNNs to graph CNNs, including contrastive gradient-based saliency maps, CAM, Grad-CAM, and Excitation Backpropagation (EB) [275].

Graphs are used within CNNs to convolve over node and edge structures to find non-Euclidean patterns. EB starts from the class output layer and traces back to the input, to show feature influence, this is different to backpropagation which looks at error propagation backwards through the layers. In their work, Grad-CAM gave the highest comparative distinctness between classes, whilst EB gave higher sparsity, and contrastive gradient-based saliency did not provide any distinction [275]. Sparsity in large graphs were deemed as more important than contrastivity for easier interpretation.

Grad-CAM is one of the most common explainability techniques used for knee OA diagnosis, used in 34 of 70 explainable AI studies [103]. However, there are only a few methods to provide explainability of healthcare diagnosis using deep learning methods [277], and even fewer for graph explainability methods, despite its importance for building trust and providing transparency to clinicians.

There are two main approaches that could be taken when thinking about explaining graph-based CNN model risk prediction, the global feature influence on prediction not related to a specific patient could be obtained, forming a kind of knowledge graph. For example, by taking the filters from a CNN layer and finding sub-graphs that represent common patterns in patient pathways which lead to hip replacement risk, a more global approach as it does not provide individual-based risk influence. Alternatively, patterns specific to a patient's pathway and their outcome that lead to a prediction risk score could be found, by extracting feature maps from models during inference on patient's EHRs, such as using CAM methods.

7.3 Methodology

7.3.1 Literature Review

A literature review was performed to find research that answered the question: How is explainable AI being used, with graph models on EHRs, to help clinicians understand what features models use to predict health outcome risk?

Google Scholar was used to search for articles using explainable graphs for health outcome prediction using the following search string:

("graph neural networks" OR GNN OR "graph models") AND ("explainability" OR "interpretability" OR "explainable AI" OR "XAI") AND ("Grad-CAM " OR "class activation map"

OR "activation mapping" OR "saliency map" OR "visuali$ation") AND ("EHR" OR "electronic health record*" OR "medical record").

The titles and abstracts were screened first, then full-texts looking for papers that met the research question criteria. Papers were included if they were written in English, published after 2010 and used graph-based models to predict health outcomes using EHRs. Grey literature and literature reviews are included in this search to get a wide overview of current methodologies being explored. Relevant papers from reviews were extracted, using the snowballing technique. A summary of the current existing literature in this area is provided in the Results (Section 7.4) of this chapter.

7.3.2 Data

The same ReseachOne dataset is used in this chapter as was used in the two previous chapters, comprising clinical data from 151,565 patients aged 40-75, with their first record of joint pain clinically coded between April 1st, 1999, and March 31st, 2014 [13].

Patients had at least two primary care visits to construct a temporal graph of their EHR. To mitigate the influence of age, sex, and socioeconomic status (IMD) on joint replacement, one-to-one matching was performed for these variables in the training dataset of each hip replacement patient. EHRs from five years prior to the hip replacement were used as the model input. Females comprised 63.6% (n = 6,664) of the cohort. The average age at the study's end (31 March 2013) was 76.8 ± 8.8 years, and the average age at hip replacement was 70.5 ± 8.5 years. Each patient had an average of 20.8 ± 42.1 (range = 707) clinical codes recorded in their EHR history.

7.3.3 Model

This chapter focuses on assessing the explainability methods on only one of the TG-CNN models. The TG-CNN model, as shown in Chapter 6, predicted hip replacement risk five years in advance with an AUROC 96.7% and a mean AUPRC of 0.879 (95% CI: 0.833-0.924) on unseen data. Unlike typical GNNs which have a node attribution matrix, a degree matrix and an adjacency matrix, the TG-CNN model uses a single 3D tensor per patient $G(i, j, k) = exp(-\gamma t_k)$. Where $i, j \in (1, .., n)$ are the graph nodes and t_k is the elapsed time between healthcare visits for the kth edge. This intuitive shape enables simple execution of CNN explainability methods,

assuming that the filters convolve over an entire space of i, j for each k. The trained TG-CNN hip replacement risk five years in advance model from the previous chapter (Chapter 6) was used with the data characteristics as described in Section 7.3.2 and Table 6.2.

The TG-CNN model architecture is as follows. The 3D CNN layer processes the $518 \times 518 \times 200$ EHR graph representation input, extracting EHR clinical code sequences and elapsed time features. The output is flattened, batch-normalised, and passed through a ReLU activation before entering an LSTM. Dropout is applied to the LSTM output, which then goes through two FCLs, followed by a ReLU before a final FCL, ReLU, and output layer. A sigmoid function outputs a binary classification result and probability. The multi-stream architecture was used (as described in Chapter 4), with one stream analysing short-term patterns and the other long-term patterns, using different filter strides (one for the original and two for the secondary stream). Outputs from both streams are concatenated after the FCL. An overview of the data to prediction to explanation process is shown in Figure 7.1.

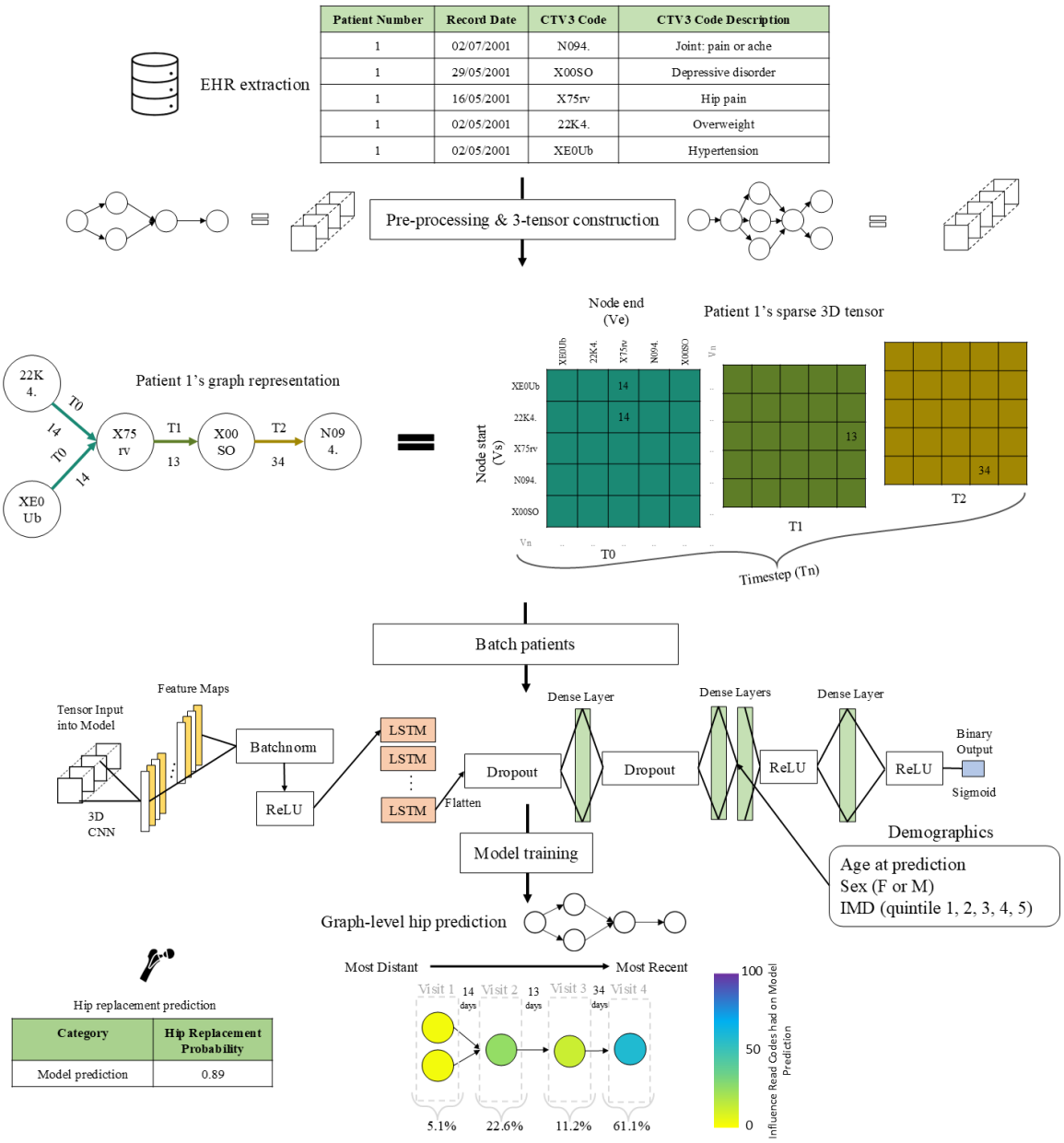


Figure 7.1: Process from a fictitious patient EHR example to graph representation, basic model architecture (just one stream is shown) and prediction outcome with explainability. A patient with five clinical codes recorded in their EHR over four primary care visits, predicted to need a hip replacement in five years time with a 0.89 probability.

7.3.4 Maximum Activation Difference

The TG-CNN model creates 1D timeseries feature maps out from its 3D CNN layer. Let A_m be one of the m feature maps from the 3D CNN layer, where the difference in maximum activation z between the two classes is calculated as follows:

$$\bar{z}_{m,c} = \frac{1}{|S|} \sum_{i \in S} max(z_i) \tag{7.1}$$

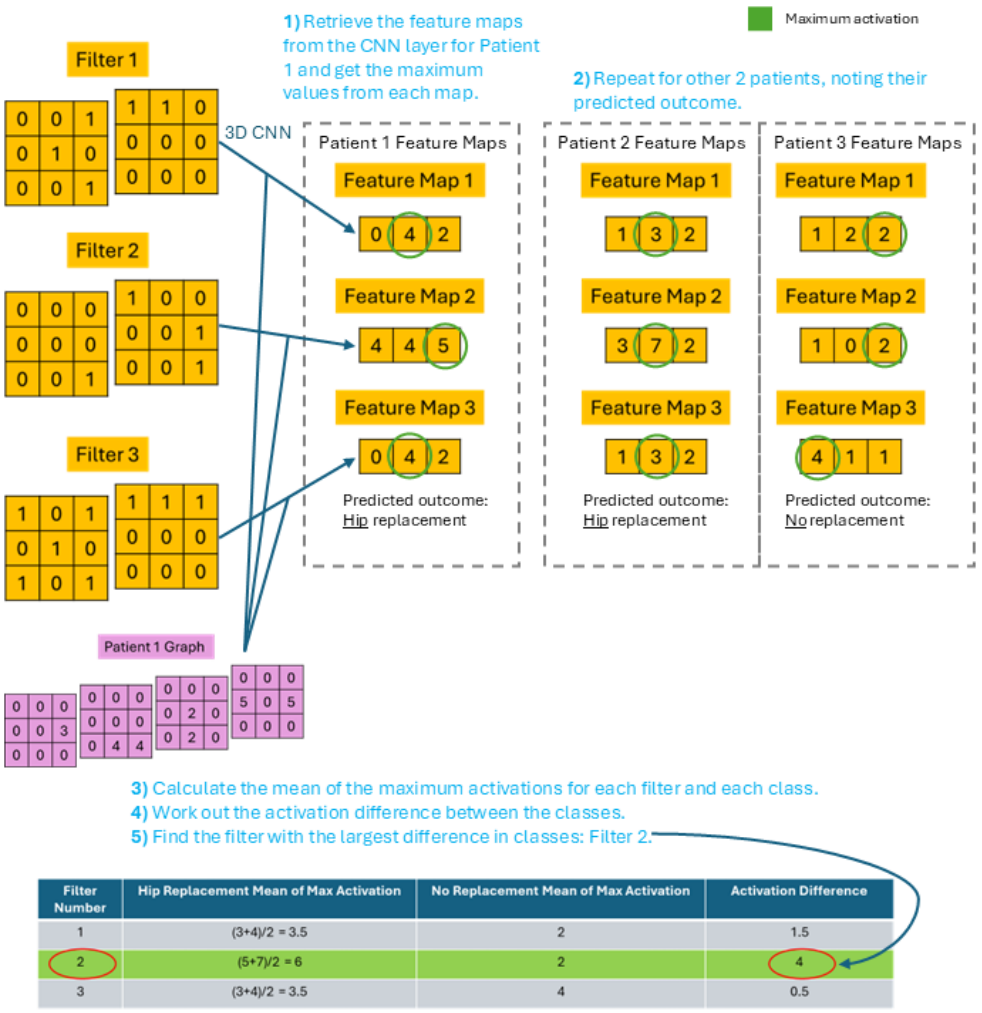
where i is a patient, z are the activation values from each feature map, m is the feature map number, and $S = (A_{m,i}$ where $y_i = c)$ is the set of feature maps A_m for the patients in class c.

The absolute difference in mean activations between the classes for each feature map is:

$$\Delta A^m = |\bar{z}_{m,c_1} - \bar{z}_{m,c_2}| \tag{7.2}$$

where c_1 is class 1 (hip replacement received) and c_2 is class 2 (no hip replacement received).

A visual representation of these equations with an example is provided in Figure 7.2.



Filter Number	Hip Replacement Mean of Max Activation	No Replacement Mean of Max Activation	Activation Difference
1	(3+4)/2 = 3.5	2	1.5
2	(5+7)/2 = 6	2	4
3	(3+4)/2 = 3.5	4	0.5

Figure 7.2: How to find the filter with the largest activation difference between the two classes for three example filters and three example patients.

7.3.5 Explainability Methods

Grad-CAM

The Grad-CAM methodology [278] is adapted for 3D graph representations. Typically, Grad-CAM highlights local features in images via heatmaps. Instead in this work activations along the axis k are given, showing which time steps were activated. Given activated time steps, primary care visit contribution towards the prediction can be observed.

The optimised weights from the trained TG-CNN model are saved and then loaded into a Grad-CAM model. A patient's graph, $G_p \in \mathbb{R}^{N \times N \times T_2}$, where N is the number of clinical codes and T_2 is the number of visits, is passed into the Grad-CAM model and the gradients are retrieved to calculate the localisation map:

The size T_1 of each 1D CNN feature map is:

$$T_1 = \frac{T_2 - filtersize}{stride} + 1 \tag{7.3}$$

Grad-CAM computes the gradient of the output Y with respect to the feature map A^m: $\frac{\partial Y}{\partial A_i^m}$ where Y is the score before the sigmoid layer and A_i^m denotes the i-th element of the 1D feature map for the m-th filter.

It then averages the gradients over the 1D feature map to get the weight α_m:

$$\alpha_m = \frac{1}{m_{size}} \sum_i \frac{\partial Y}{\partial A_i^m} \tag{7.4}$$

where α_m is the weight for the m-th filter.

Finally, the weighted sum of the feature maps using the weights α_m are calculated and passed through a ReLU:

$$L_{\text{Grad-cam}} = \text{ReLU}\left(\sum_m \alpha_m A^m\right) \tag{7.5}$$

The resulting 1D localisation map $L_{\text{Grad-CAM}}$ can be used to understand which parts of the input sequence are most important for the classification decision.

An average weight for each time step is calculated using the time steps that correspond to the

filter window. To assign the $L_{\text{Grad-CAM}}$ values to the original patient graph G_p, the value is spread equally across the d time steps in G_p which form each element of V. Where $V \in \mathbb{R}^{T_2}$ is the vector of $L_{\text{Grad-CAM}}$ values at each time step k.

Let $w \in \mathbb{R}^{T_2}$ be the weight at each time step in G_p.

Let X_m be the indicator function and v_m is each value within V.

$$X_m(i) = \begin{cases} 1 & \text{if } k_i \text{ in } G_p \text{ contributes to } v_m, \\ 0 & \text{otherwise} \end{cases}$$

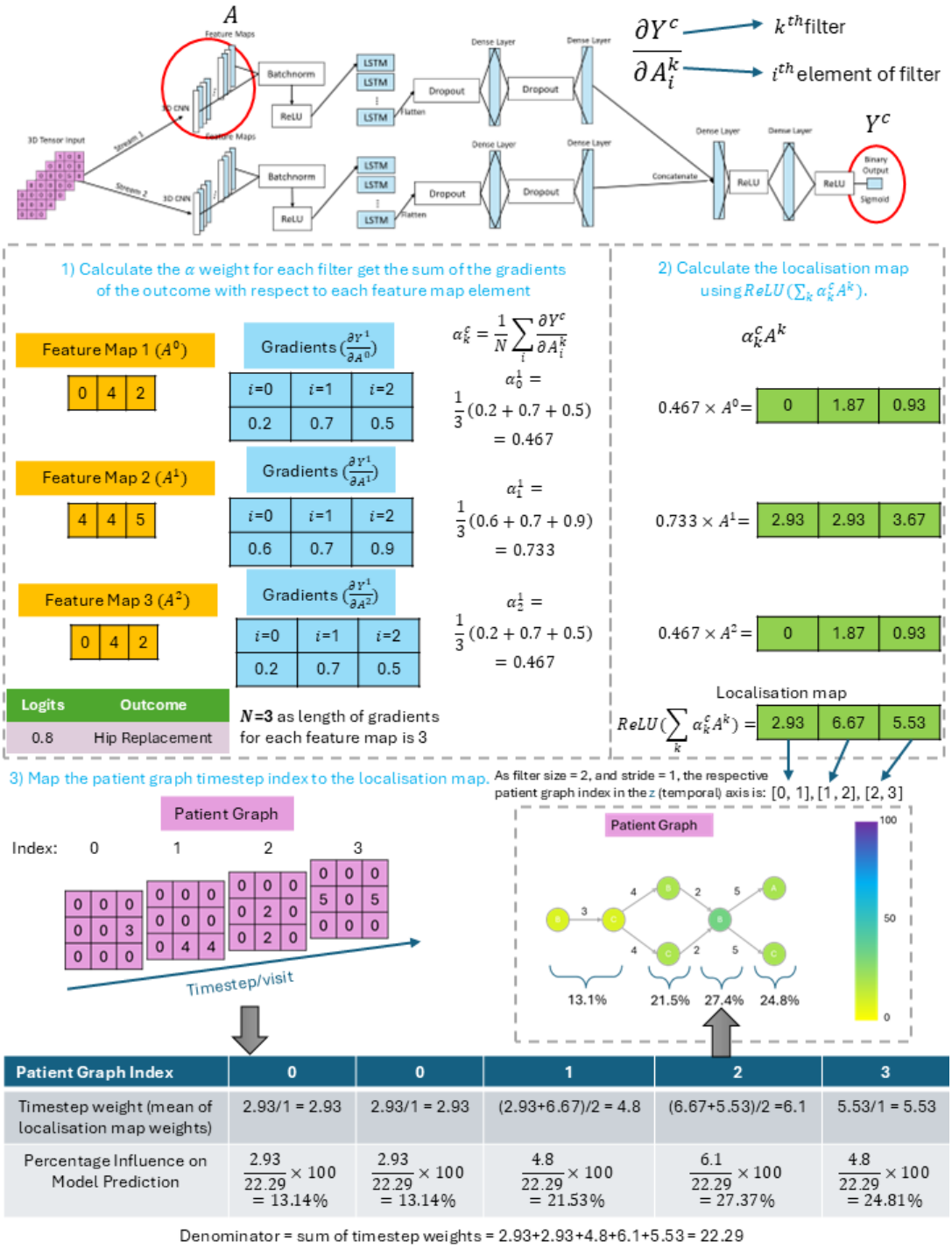
Then

$$w_i = \frac{1}{d} \sum_{m=1}^{T_1} X_m(i) v_m \tag{7.6}$$

See Figure 7.3 for a visualisation of this calculation applied to an example patient's graph.

Due to using a ReLU function, Grad-CAM only pays attention to areas of the graph which increase patient risk. To also consider events which lower risk, a modified version called Grad-CAM (abs) is used by editing Equation 7.5 to:

$$L_{\text{Grad-CAM}} = \text{abs}\left(\sum_m \alpha_m A^m\right) \tag{7.7}$$



Patient Graph Index	0	0	1	2	3
Timestep weight (mean of localisation map weights)	2.93/1 = 2.93	2.93/1 = 2.93	(2.93+6.67)/2 = 4.8	(6.67+5.53)/2 =6.1	5.53/1 = 5.53
Percentage Influence on Model Prediction	$\frac{2.93}{22.29} \times 100 = 13.14\%$	$\frac{2.93}{22.29} \times 100 = 13.14\%$	$\frac{4.8}{22.29} \times 100 = 21.53\%$	$\frac{6.1}{22.29} \times 100 = 27.37\%$	$\frac{4.8}{22.29} \times 100 = 24.81\%$

Denominator = sum of timestep weights = 2.93+2.93+4.8+6.1+5.53 = 22.29

Figure 7.3: Grad-CAM (ReLU) methodology visualised: computing the gradient of the outcome class with respect to the feature/activation map by gradient tracking and then applying the localisation map back onto the patient graph.

fm-act Graphs

The feature maps from the CNN layer can indicate the concepts learned by the TG-CNN model. The fm-act methodology is proposed as below:

Extract the feature maps: $A = \{A_1, A_2, \ldots, A_n\}$ from the 3D CNN layer of the TG-CNN model.

Let A^* denote one of the following summaries: the feature map with the strongest class differentiation $\max_m \Delta A_m$, mean(A), or median(A).

Map the feature map weights to the time steps and get an average of the weights for each sliding window recurrence on each time step:

$$W_k = \frac{1}{|\mathcal{W}_k|} \sum_{(i) \in \mathcal{W}_k} A_i^* \tag{7.8}$$

where $\mathcal{W}_k$ is the set of all sliding windows that include time step k.

Normalise the weights to get the percentage influence of each time step: $W_{\%}^k = \frac{W_k}{\sum_{k'} W_{k'}} \times 100$.

A visualisation of these steps for a small example patient graph can be seen in Figure 7.4.

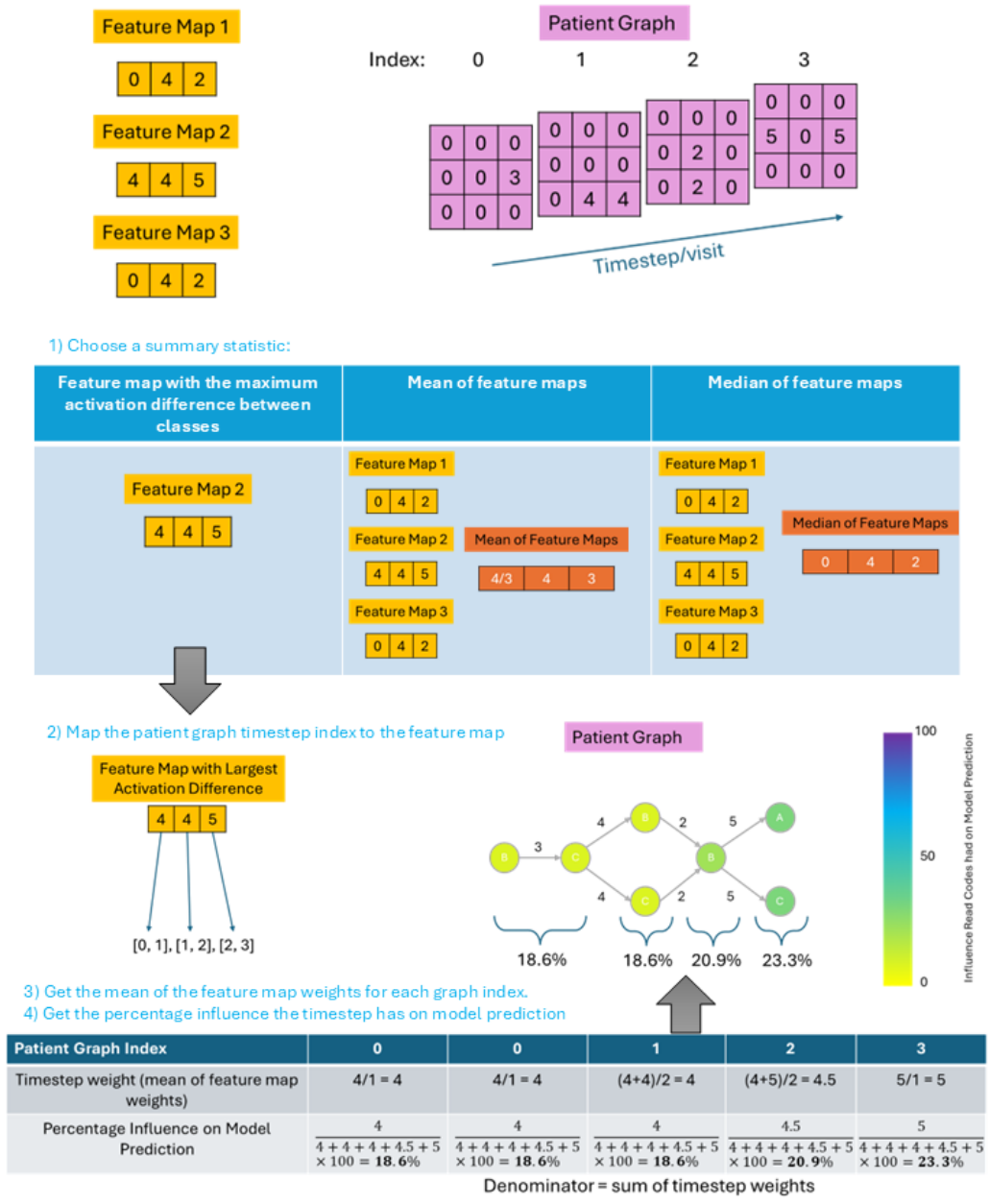


Figure 7.4: How to create feature map graphs showing timestep/visit influence on prediction decision.

edge-act Graphs

As another way of obtaining feature importance is using the filters from the CNN layer. The filters can be observed and used alongside the original input graph to find edge importance.

Extract the filters: $F = f_1, f_2, \ldots, f_n$ from the 3D CNN layer of the TG-CNN model.

Let f^* denote one of the following: the filter f^m with the feature map with the strongest class differentiation $\max_m \Delta A_m$, mean(F), or median(F).

Compute the edge-act via a 3D sliding window:

$$E_{i,j} = \frac{1}{|\mathcal{W}i,j,k|} \sum_{(a,b,c)\in\mathcal{W}i,j,k} \left(G_{a,b,c} \odot f^*_{a,b,c}\right) \tag{7.9}$$

where $\mathcal{W}_{i,j,k}$ is the 3D sliding window over the input graph G for each filter position (i, j, k) and $\odot$ is element-wise multiplication.

Normalise the weights to get percentage influence of each edge: $W_{\%} = \frac{W_{ij}}{\sum_{i,j} W_{ij}} \times 100$.

See Figure 7.5 for a visual walkthrough of these steps to an example patient graph.

7.3.6 Interactive Visualisations

Plotly (5.23.0) and NetworkX (3.2.1) were used to plot interactive graphs. A patient's individual graph is shown where the nodes are the clinical codes, stacked clinical codes occurred during the same visit, the edges show the days between visits. A patient's risk of requiring a hip replacement in five years is also provided.

In the Grad-CAM and fm-act explainability graphs, nodes with higher influence on the model prediction (larger w_i values) are coloured darker blue/purple, while those with lower influence are lighter yellow. The Viridis heatmap was used for visual accessibility. Users can hover over nodes to view clinical code descriptions and percentage influence. In the edge-act explainability graphs, edges with greater influence (larger $\mathcal{W}_{\%}$ values) are darker blue/purple, and less influential edges are lighter yellow. Hovering over edges reveals the percentage influence.

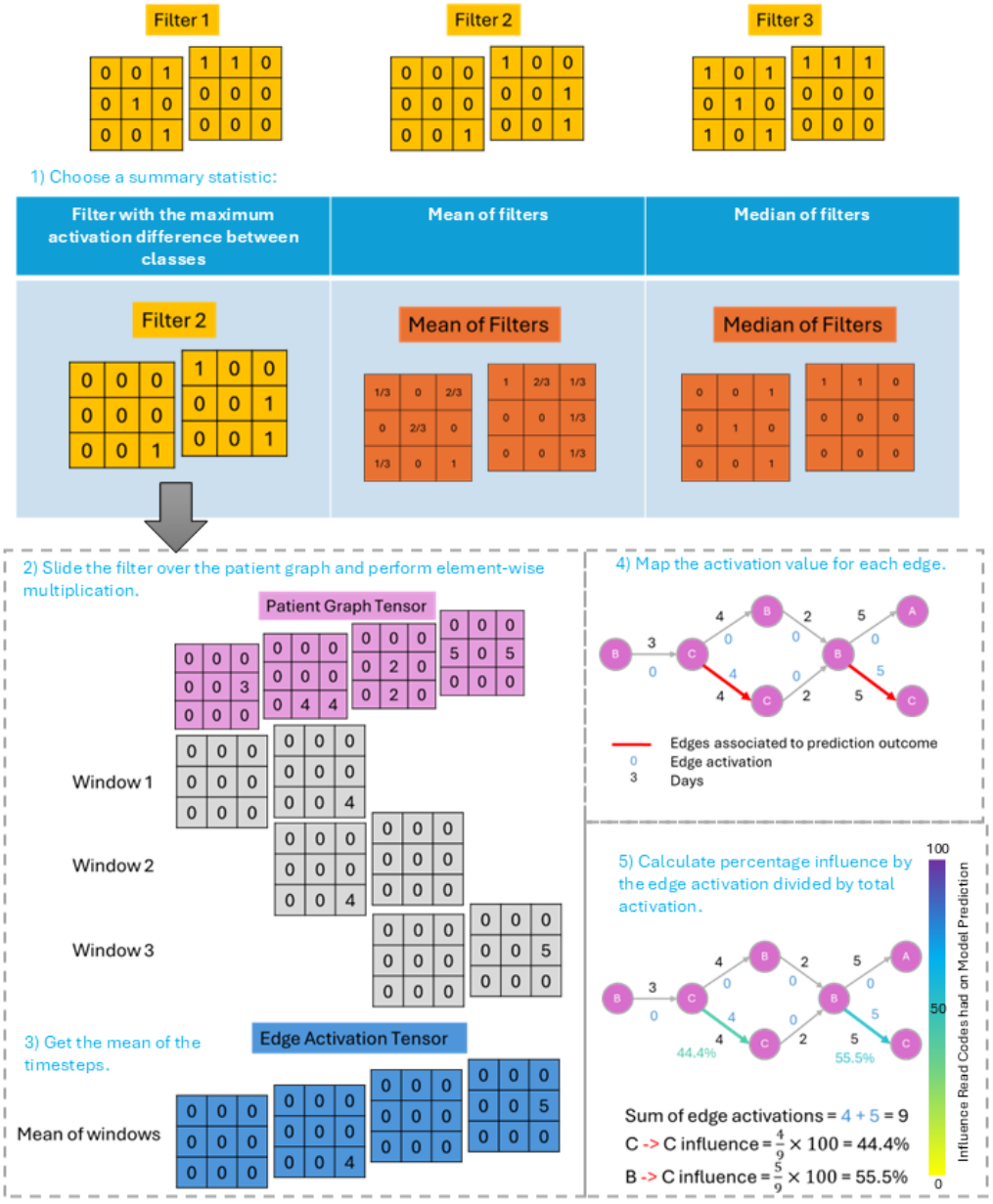


Figure 7.5: How to create edge graphs showing clinical code pair influence on prediction decision.

7.3.7 Evaluating Graph Visualisations

Saliency maps may not always be reliable for understanding model decision making as they rely on intuition, have poor falsifiability [276], and tend to represent noise rather than signals [279]. For this reason the methods are scored using the evaluation techniques below.

Sensitivity: For the Grad-CAM and fm-act methods a node with a random clinical code assignment is added to a random existing visit 10 times, then the mean L1 distance between the original and edited visit is obtained. For the edge-act model, a node's clinical code is randomly changed and the influence of the connected edge going into the node to the original edge influence is compared. Higher values determine higher methodology sensitivity.

Edge Detection Bias: The model weights were changed by randomly adding noise (with the same mean and standard deviation as observed from the trained model weights) to observe changes in the heatmaps, which are used to suggest the observed influence of edges and visits. If EDB (false saliency) is present then the heatmaps will not change or will be very similar. Ideally, the percentage influence will be dependent on the weights of the model. The mean absolute error (MAE) is calculated for the difference between the heatmap from the trained model and the heatmap from the random weighted model for each of the methods, then the mean and standard deviation of the differences across the patients is retrieved. A higher EDB value indicates false saliency is less likely.

Sparsity: Node and edge weight sparsity were calculated by binarising the heatmaps, setting values over 0 to 1 and others to 0. The sparsity for each graph was taken to be the percentage of non-zero entries. A larger sparsity value indicates more nodes or edges with no influence on model prediction [275], which may be better for larger graphs, to make visual focus more obvious.

Human Evaluations: Qualitative human evaluation studies were also carried to assess the interpretability versus truth trade-off of the four explainability methods, whilst gauging user interaction experience [234]. Seven clinicians were asked to complete a survey which showed the four graph methods visualised (a = Grad-CAM original, b = Grad-CAM-abs, c = fm-act, and d = edge-act) for 5 different patient cases. A clinical vignette is provided in the Appendix to explain the visualisation (Appendix D.1). The survey questions the clinicians were asked are

in Appendix D.1, this included free text questions and Likert scales (no statistical significance testing was carried out).

Subgraph Frequency Analysis: Once the sensitivity, EDB, and sparsity of each method were collected, the optimum method was selected and subgraph analysis was performed on it. To analyse the frequency of subgraphs across the patients, first all of the edges with a percentage influence of more than 0 (denoted as 'activated' nodes) were found. Then the collections of connected activated nodes and their connection edges were collected, repeating this for all patient graphs. The subgraph frequency was counted by prevalence per class N_+^s, N_-^s for subgraph s, then the ratio was obtained to give subgraph prevalence by class: $R_+^s = \frac{N_+^s}{N_+^s+N_-^s}$, $R_-^s = \frac{N_-^s}{N_+^s+N_-^s}$ [275]. It is important to note that the subgraphs obtained have not been established as definite patient trajectories that influence hip replacement risk, instead these subgraphs illustrate how the model infers prediction globally.

Further to this three questions were considered: 1) Are there differences in the most common subgraphs between the train and test datasets? 2) Which subgraphs are most commonly accurately and inaccurately classified? 3) Which subgraphs most commonly influence model prediction?

The correctness/ accuracy of a subgraph was calculated by comparing the patient's actual outcome to the model's predicted outcome. A predicted probability above 0.5 was interpreted as indicating the need for a replacement, while probabilities of 0.5 or below were interpreted as no replacement needed. However, the model is designed as a risk assessment tool rather than a strict binary classification system. Assigning a definitive decision based on a 51% predicted risk does not align with the intended use of the model, which evaluates likelihood rather than providing categorical outcomes.

7.4 Results

7.4.1 Literature Review

Inputting the search string into Google Scholar resulted in 111 papers being returned. After title screening 56 papers were left. After abstract screening 18 papers remained. Fifteen of these 18 papers were literature reviews, from which snowballing of references was performed finding another 23 papers, totalling 26 full-papers to screen. See Figure 7.6 for the PRISMA

flowchart [157]. In total three papers used graphs and AI with explainability methods using EHRs for healthcare prediction [280, 281, 282], all of these papers used attention mechanisms and incorporated temporality into their explainable models.

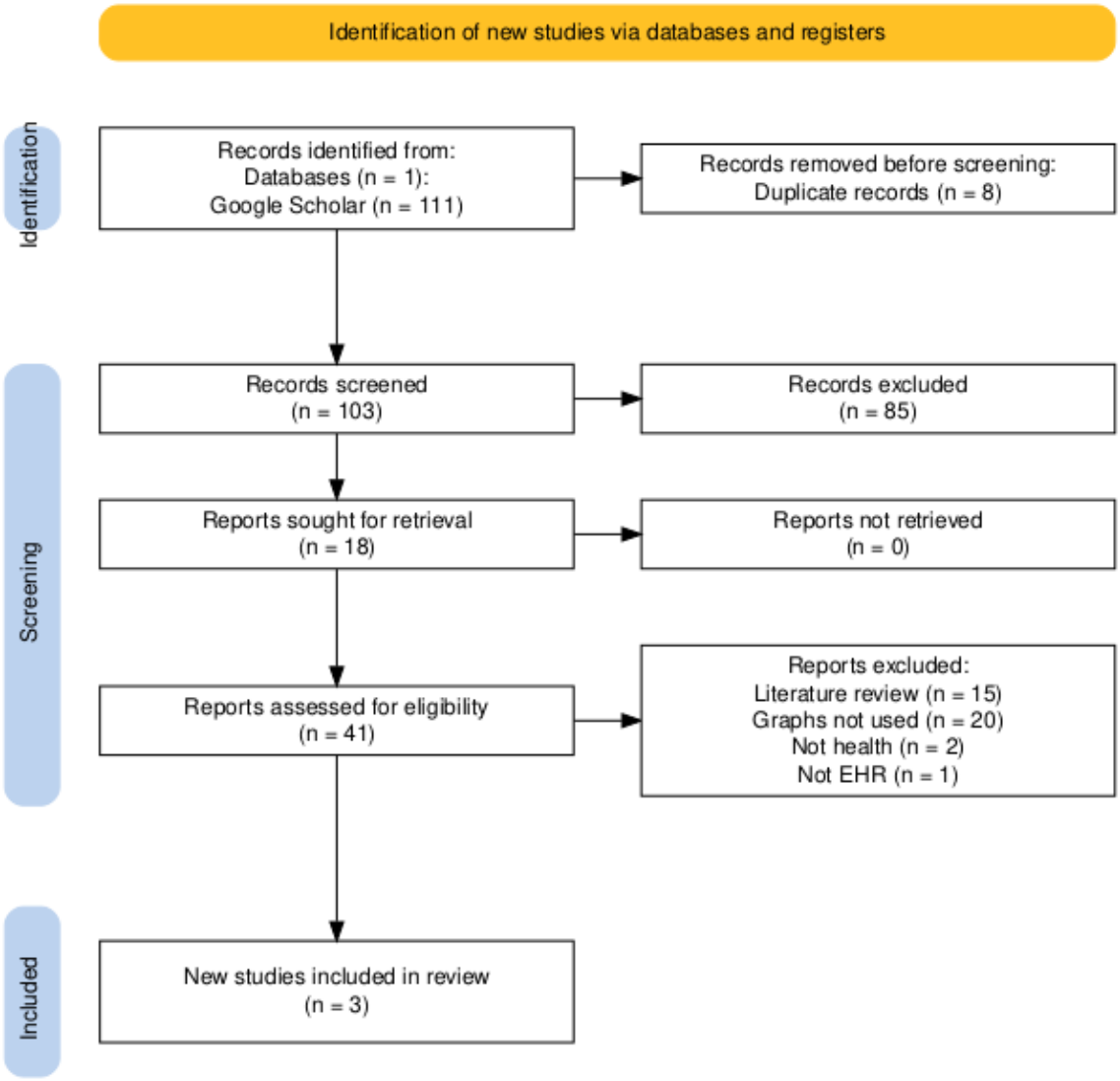


Figure 7.6: PRISMA flowchart of search for explainable graph methods using EHRs.

Su et al. 2020 presented the GATE model, a Graph-Attention Augmented Temporal Neural Network for medication recommendation, which captures relationships between symptoms, medications, and temporal changes in medical history at each admission. The graph attention mechanism enables the model to prioritise important aspects of the medical history, particularly the temporality. Sun et al. 2021a integrated medical event temporality, frequency, and attention mechanisms on to graph structures to predict outcomes such as mortality and readmission, with graph nodes representing diagnoses, symptoms, and treatments [281]. Sun et al. 2021b utilised GNNs, RNNs, and attention mechanisms to predict patient survival times [282].

7.4.2 Methodology Comparison

Figure 7.7 shows the distribution of maximum activation values from each CNN feature map for all patients. Here it can be seen that the distributions are non-Gaussian with skew. Figure 7.8 demonstrates the difference in maximum activation between the classes, this can be used to find which filter has the largest distinction in activations when predicting whether a hip replacement will be needed in five years time. Feature map 30 had the largest difference in activation between the classes, whilst feature map 1 had the smallest difference.

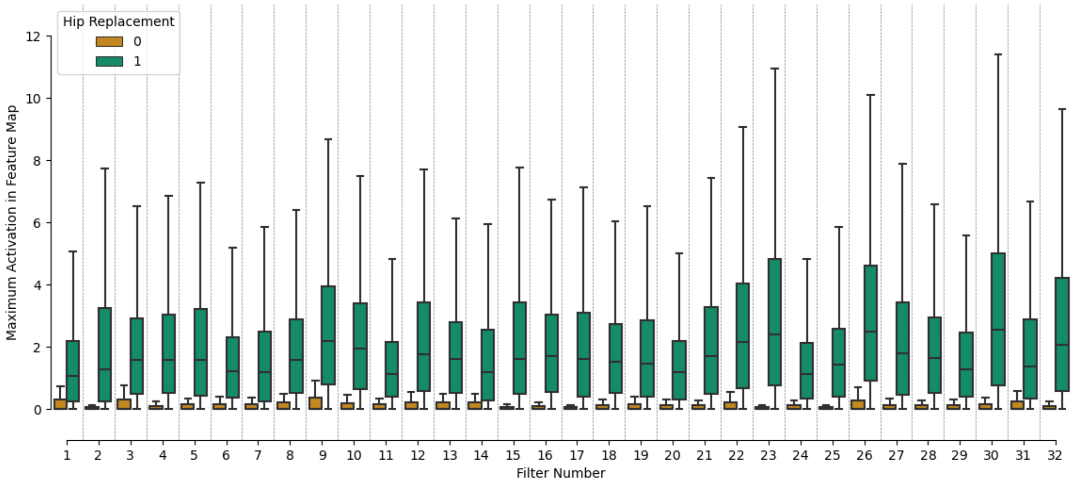


Figure 7.7: Boxplot showing maximum activation for each filters feature map, for both classes and all patients.

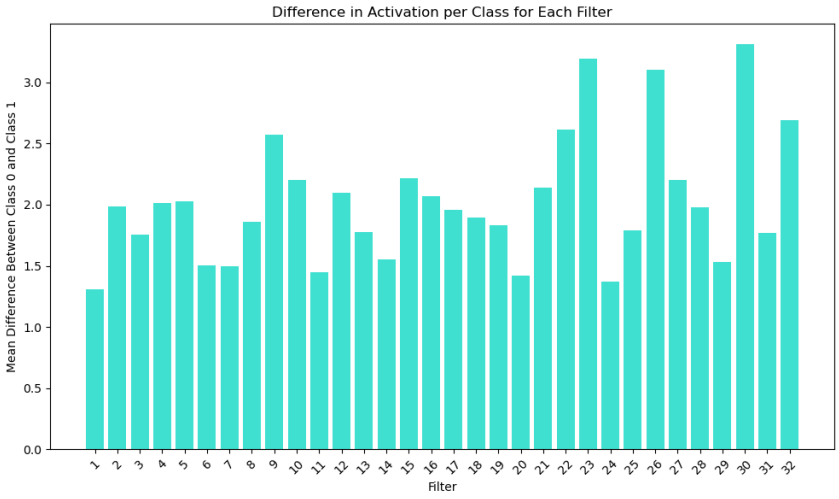


Figure 7.8: Bar chart showing the difference in maximum activation between the two classes.

Table 7.1 shows the results from evaluating the different methodologies.

The Grad-CAM models resulted in higher normalised mean gradient weight values from the more recent visits and close to 0 gradient values elsewhere, forming exponential curves, see

Table 7.1: Evaluation results mean (standard deviation). Edge detection bias (EDB), mean absolute error (MAE).

	Sensitivity	EDB MAE	Sparsity
grad-cam			
ReLU	4.59(5.94)	4.40(10.92)	0.30(0.43)
Abs	5.80(5.90)	2.16(4.53)	0.00(0.00)
fm-act			
Mean	6.05(6.19)	2.25(1.85)	0.00(0.00)
Median	5.96(5.75)	1.65(1.32)	0.00(0.00)
Max	6.18(5.95)	0.03(0.03)	0.00(0.00)
edge-act			
Mean	25.00(23.97)	9.82(21.07)	0.53(0.32)
Median	23.64(23.26)	4.40(4.56)	0.55(0.31)
Max	23.78(23.20)	15.89(26.77)	0.51(0.33)

Figure 7.9. More varied normalised mean activation values appeared throughout the feature map, whereas Grad-CAM gradient weights produced smoother curves.

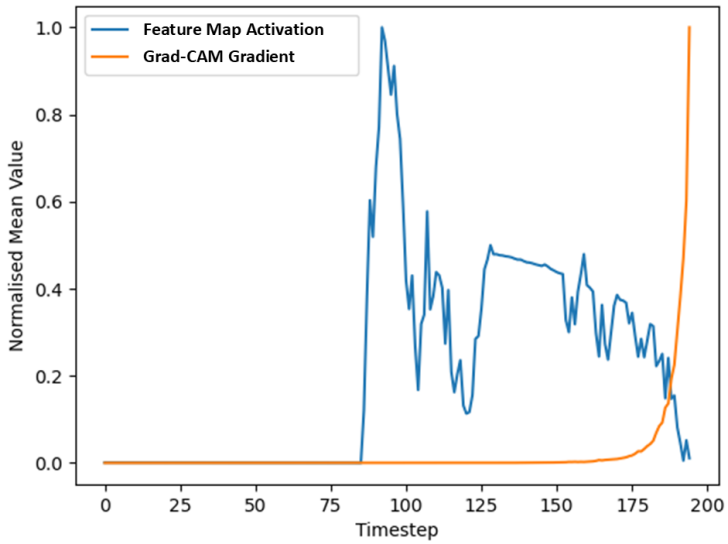


Figure 7.9: Normalised mean gradient and feature map activation values from one patient.

Figures 7.10 - 7.13 show the heatmaps for one patient with the four explainable methodologies, this indicates how the influence of each visit or edge fluctuates depending on the filter and the method.

7.4.3 Interactive Visualisations

Figure 7.14 shows the graphs used to show an example patient pathway and influence healthcare records have on model prediction. When these models are interacted with as HTML files users

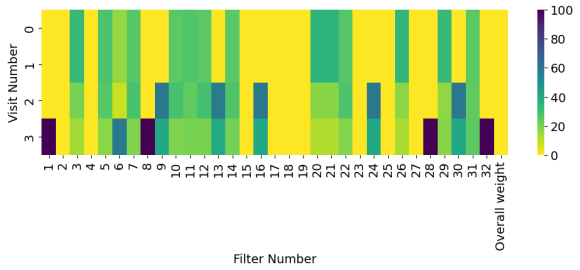


Figure 7.10: Heatmap of percentage influence using Grad-CAM (ReLU).

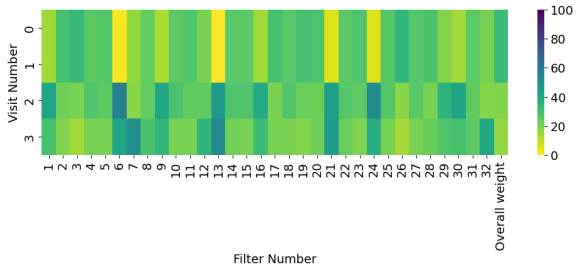


Figure 7.11: Heatmap of percentage influence using Grad-CAM (abs).

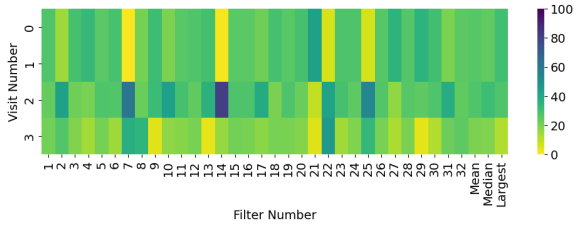


Figure 7.12: Heatmap of percentage influence using fm-act.

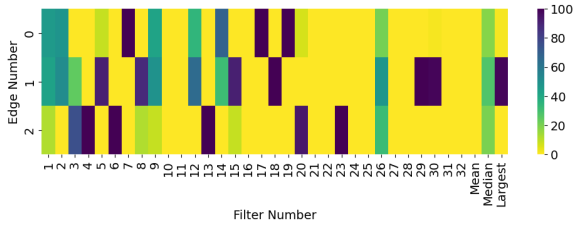


Figure 7.13: Heatmap of percentage influence using edge-act.

can hover over nodes (Grad-CAM and fm-act graphs) and edges (edge-act graphs) to show the percentage influence to the model decision and the clinical code descriptions[1]. The examples provided have short patient history for simplicity when presenting as a static PNG file, however the HTML plotly plots enable the users to navigate the graphs.

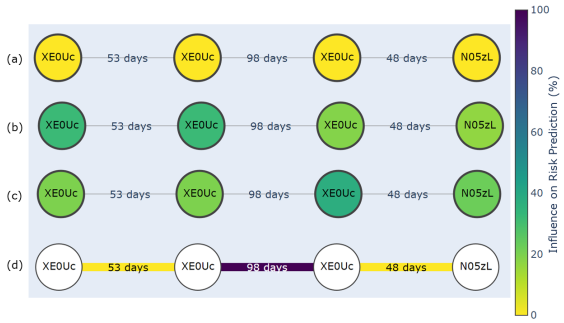


Figure 7.14: Percentage influence on features using 4 methods: (a) Grad-CAM (ReLU), (b) Grad-CAM (abs), (c) max fm-act, and (d) median edge-act. Here the patient's predicted risk was 3.61% and they did not receive a hip replacement. Clinical code descriptions: XE0Uc = Essential Hypertension and N05zL = Osteoarthritis of knee.

7.4.4 Clinical Feedback

Some patients had long EHR histories which led to complex graph visuals, whilst others were simpler and easier to visualise without zooming or panning to specific areas of the graph.

[1]The reader can interact with example graphs themselves here: *https://zoehancox.github.io/graph-survey/index.html*

Overall, the max fm-act method (c) was voted as the easiest graph to interpret (n = 15/35), and Grad-CAM (abs) (b) the hardest (n = 1/35), see Figure 7.15. There were varying opinions on the effectiveness of methods in highlighting key factors influencing model predictions across different patient graphs, with the two longest patient graphs having the worst feedback (Figure 7.16). Overall, satisfaction with the methods decreased as the length of the patient history increased (Figures 7.16 and 7.17). Patient graphs 2 and 3 had the most agreement from clinicians for expected trajectory alignment, there was 57% agreement in all the patient graphs that trajectories met expectations, however some methods were selected as not aligning with expectations (Figure 7.17). When clinicians were asked as to what extent did the graphs support understanding the model's decision-making process, one said it 'Greatly Supported', two said it 'Moderately Supported', three said it 'Slightly Supported', and one scored it as 'Neutral'. Three clinicians felt that these graph visualisations had 'Neutral' input for aiding decision-making in a clinical-setting, one thought they were 'Very Useful', one thought they were 'Somewhat Useful', and two clinicians felt they were 'Somewhat Useless'. These graphs appeared useful to demonstrate model decision-making, but were less helpful for aiding clinical decision-making.

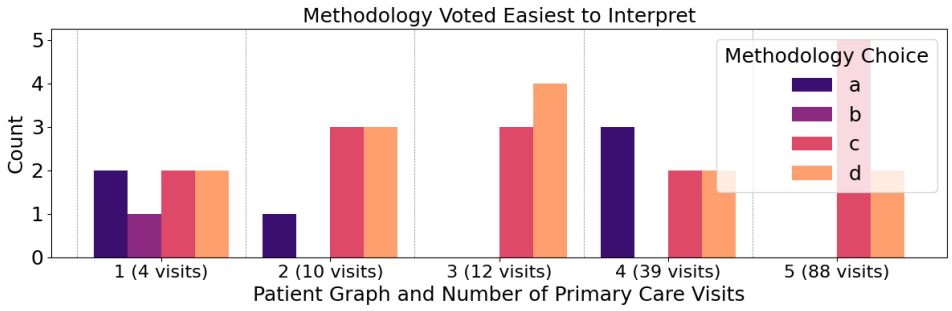


Figure 7.15: Which method was determined the easiest to visually interpret.

The survey feedback highlighted that method c stands out for its subtle differences and ease of engagement, with better colour discrimination between nodes. Methods a-c more effectively addressed model complexity. However, there was a general consensus that the colour scale should be given more emphasis across all methods. Method d received mixed reviews: while one clinician found it visually unappealing, another preferred it over methods a and b for its visual clarity. Additionally, there was some confusion from a clinician regarding the model's decision process, specifically questioning the connection between hypertension and hip replacement.

The survey results suggest that the current graph layout is too crowded and detailed for GPs

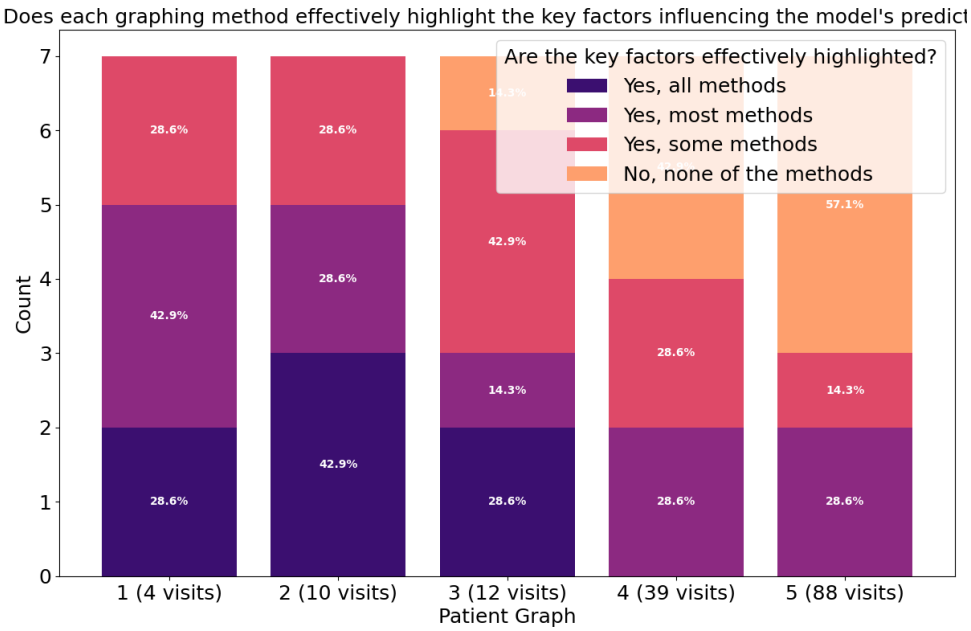


Figure 7.16: Clinical opinion on whether key factor influence is highlighted.

to effectively use within the time constraints of a primary care appointment. A visual summary of the most influential factors is recommended to facilitate quicker decision-making and patient communication. While detailed graphs are valuable for understanding model decision-making, a focus on summary risk prediction scores is deemed more practical. Simplifying the colour coding could improve clarity and distinction where percentages are similar, though this may limit the ability to compare multiple patients visually.

7.4.5 Subgraph Frequency Analysis

Subgraphs were collected using the median edge-act method as outlined in Section 7.3.7 as this method had the highest sparsity. A total of 13,383 subgraphs were identified. 10,594 subgraphs only appeared once (79.2%). The most frequent subgraph for patients with high hip replacement risk was NOA $\rightarrow$ NSAIDs prescription, with many of the most frequent subgraphs containing prescriptions. See Figure 7.18 for the 10 most frequent subgraphs influencing model prediction for each class.

Table 7.2 shows the number of subgraphs present from the models from each subset of data. Note that this is not the total number of possible subgraphs, but instead the number of subgraphs that occur from this methodology. The one year in advance knee replacement prediction model had the most subgraphs, whilst the five years in advance hip replacement model had the least subgraphs. The hip one year in advance model had the most overlap in subgraphs between the

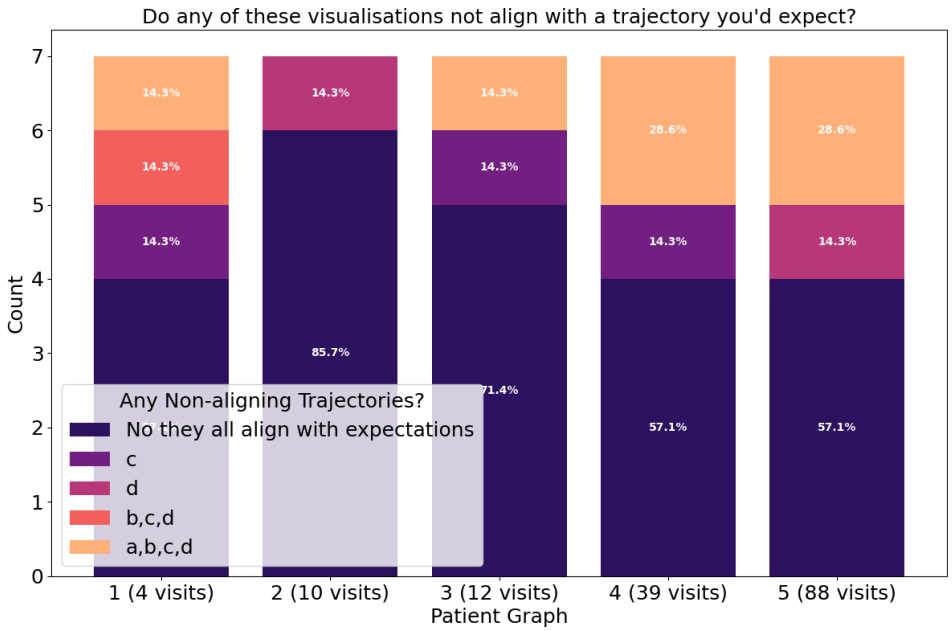


Figure 7.17: Clinical opinion on trajectory alignment.

training and testing datasets (10.77%), whilst the knee five years in advance model had the least crossover (8.08%). The hip one and five years in advance models have 4.13% similarity in unique subgraphs, whilst the knee one and five years in advance models have 3.30% overlap in subgraphs. The hip and knee models have 5.35% identical subgraphs.

Table 7.2: Number of subgraphs produced in each dataset alongside longest subgraph length and overlap (%) of subgraphs present in both groups.

Data	**Training Data**	**Testing Data**	**Total Sub-graphs**	**Overlap (%)**	**Max Number of Visits**	**Max Number of Nodes**
Hip (1 year)	33,181	59,532	92,713	10.77	87	360
Hip (5 years)	13,383	52,728	66,111	7.97	38	76
Knee (1 year)	42,918	79,750	122,668	9.67	107	202
Knee (5 years)	14,108	53,349	67,457	8.08	34	76

Figure 7.19 shows the 10 most frequently presented subgraphs in the hip five years in advance cohort with predictions in the positive and negative classes. Both training and testing subgraphs show the high frequency of prescription records, weight, and smoking status in patients with higher hip replacement risks. For the most frequent subgraphs for low-risk of hip replacement, prescriptions were less prevalent, whilst HTN appeared more, and BMI records occurred more frequently.

Figure 7.20 shows the 10 most frequent subgraphs that were found in models which were cor-

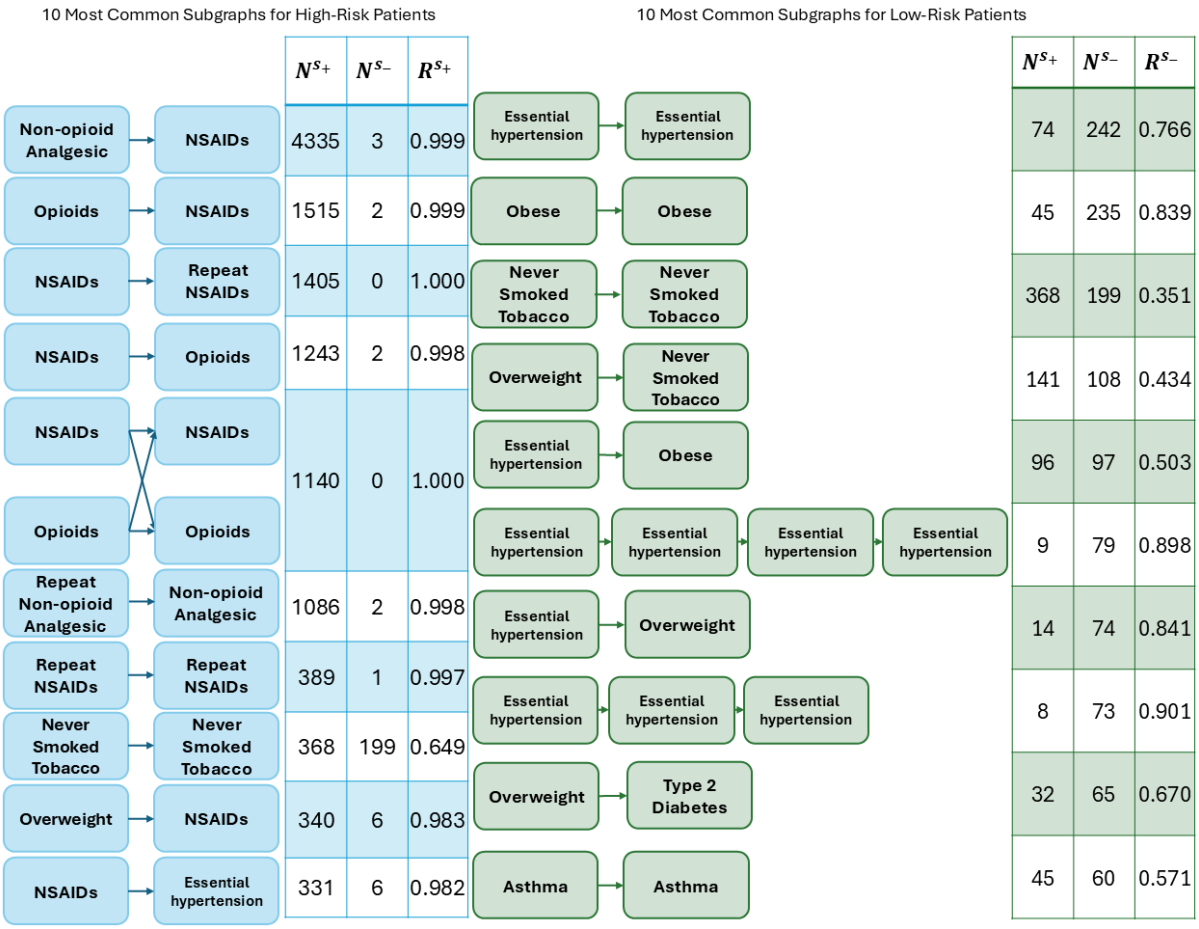


Figure 7.18: The 10 most frequent subgraphs that influence model prediction for each class.

rectly classified. It should be noted that other subgraphs may be present in individual EHRs that are more informative to prediction than these subgraphs. For example, Angina pectoris → HTN does not mean that someone will have a higher hip replacement risk, the full EHR may contain other key indicators such as hip pain or OA. It is also worth emphasising again that these subgraphs do not show what causes replacement risk, these subgraphs show pathways that the model used for prediction which decisions may not align with clinical expertise.

Figure 7.21 shows the 10 most frequent subgraphs that were found in models which were incorrectly classified.

The subgraph with the highest influence from each patient was selected see Figure 7.22, to determine which subgraphs are most commonly influential for model classification decision, for each outcome type.

When selecting only the most influential subgraphs from each patient in the hip five years in advance dataset, there were only 3,438 different subgraphs in those predicted to require a

future hip replacement, with the largest subgraph having 76 nodes. There were a maximum of 36 visits in this cohort. For those predicted to not need a hip replacement in five years time, there were 9,762 subgraphs (corresponding to the negative class), with the most influence on outcome decision. There were 60 nodes in the subgraph with the largest number of nodes, and 29 visits in the subgraph with the largest number of visits. Figure 7.23 shows the subgraphs most frequently given the highest influence for each patient, per class outcome.

10 Most Common Subgraphs for High-Risk Patients

Hip 5 Years in Advance Training Data Subgraphs				N^{s+}	N^{s-}
NOA →	NSAIDs			4335	3
Opioids	NSAIDs			1515	2
NSAIDs	rNSAIDs			1405	0
NSAIDs	Opioids			1243	2
NSAIDs	NSAIDS			1140	0
Opioids	Opioids				
rNOA	NOA			1086	2
rNSAIDs	rNSAIDs			389	1
Never smoked	Never smoked			368	199
Overweig ht	NSAIDs			340	6
NSAIDs	HTN			331	6

Hip 5 Years in Advance Testing Data Subgraphs				N^{s+}	N^{s-}
NOA →	NSAIDs			19687	11
rNOA	NOA			10893	2
Opioids	NSAIDs			9325	5
NSAIDs	Opioids			6409	8
NSAIDs	NSAIDS			5857	2
Opioids	Opioids				
NSAIDs	rNSAIDs			5669	6
Ex-smoked	Ex-smoker			3879	93
Never smoked	Never smoked			3268	90
Never smoked	NOA			3043	2
NOA	Never smoked			2882	0

10 Most Common Subgraphs for Low-Risk Patients

Hip 5 Years in Advance Training Data Subgraphs				N^{s+}	N^{s-}
HTN →	HTN			74	242
Obese	Obese			45	235
Never smoked	Never smoked			368	199
Overweig ht	Never smoked			141	108
HTN	Obese			96	97
HTN	HTN	HTN	HTN	9	79
HTN	Overweig ht			14	74
HTN	HTN	HTN		8	73
Overweig ht	Type 2 Diabetes			32	65
Asthma	Asthma			45	60

Hip 5 Years in Advance Testing Data Subgraphs				N^{s+}	N^{s-}
Ex-smoker →	Ex-smoker			3879	93
Never smoked	Never smoked			3268	90
Obese	Never smoked			1650	47
Overweig ht	Never smoked			1127	34
Obese	Obese			234	29
HTN	HTN			70	24
HTN	Obese			511	23
Type 2 diabetes	Obese			367	23
Never smoked	Overweig ht			652	19
NOA	Obese			2322	17

Figure 7.19: Hip five year in advance model training data vs testing data subgraphs. 10 most frequent subgraphs for positive and negative classes. Hypertension (HTN), Non-steroid anti-inflammatory drugs (NSAIDs), repeat prescription of NSAIDs (rNSAIDs).

Hip 5 Years in Advance Subgraphs			N^{s+}	N^{s-}	Predicted Correctly
Angina pectoris →	HTN		**8**	0	8
HTN disease	Hip pain		**2**	**5**	7
Asthma	Angina pectoris		**7**	0	7
	HTN				
IHD	Asthma				
COAD	Chronic depression		**6**	0	6
IHD	Hip pain		**2**	**4**	6
HTN	OA		**1**	**5**	6
HTN	NOA	Never smoked	**6**	0	6
Malignant prostate tumour		Obese			
		HTN			
Overweight	Underweight		**2**	**4**	6
COAD	Acute exacerbation of COAD		**4**	**1**	5
Obese	Obese		0	**5**	5
	DM				

Figure 7.20: Most frequent subgraphs accurately classified for the hip five year in advance model. Ischaemic heart disease (IHD), Chronic obstructive airways disease (COAD), Osteoarthritis (OA), Diabetes mellitus (DM), Non opioid analgesic (NOA).

Hip 5 Years in Advance Subgraphs			N^{s+}	N^{s-}	Predicted Correctly
Ex smoker →	Ex cigarette smoker		72	0	0
NOA	Opioids		70	1	0
NOA	Ex cigarette smoker	NOA	67	0	0
NOA	Obese Ex cigarette smoker		50	0	0
Overweight	NOA	Never smoked	45	0	0
Opioids	Healthy weight Smoker Cigarette consumption		43	0	0
Ex cigarette smoker	Opioids		41	0	0
Opioids	Obese Smoker	Opioids	39	0	0
Depressive disorder	NSAIDs		39	0	0
NSAIDs	Smoker Cigarette consumption	NSAIDs	38	0	0

Figure 7.21: Most frequent subgraphs incorrectly classified for the hip five year in advance model.

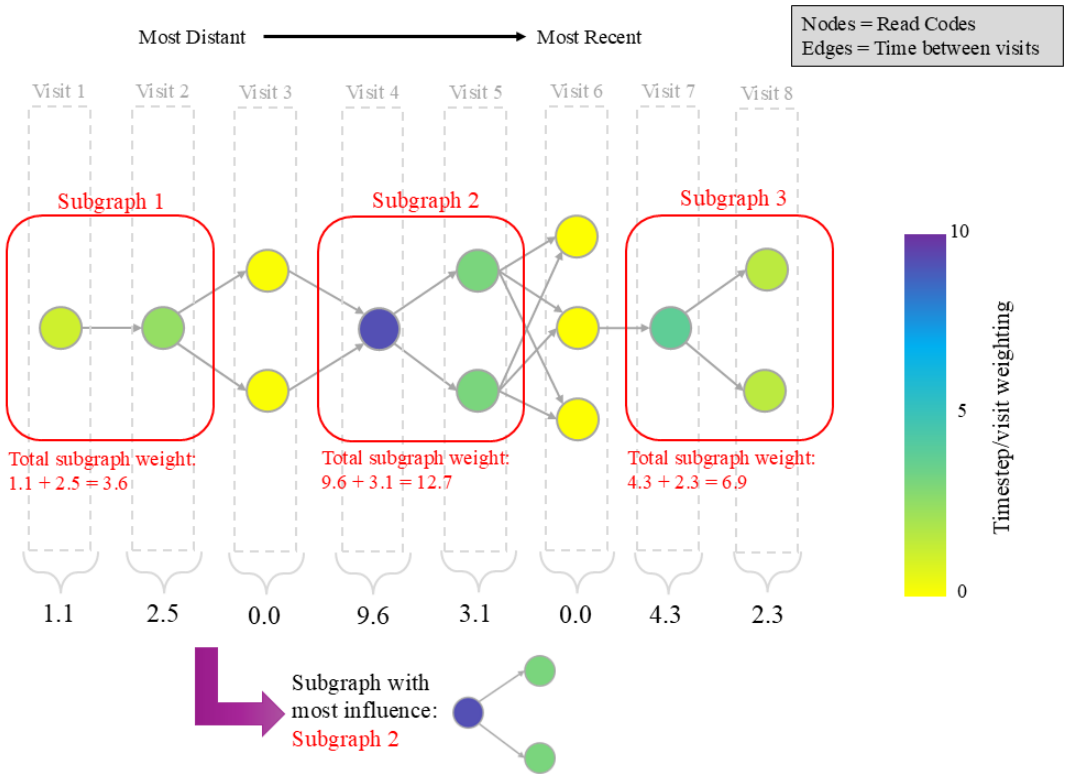


Figure 7.22: How the most influential subgraph was selected from each patient. Taking the subgroup (section of nodes with >0 weight) with the highest weighting.

Positive Class

10 Most Frequent Influential Subgraphs for Patients Recommended to Have a Hip Replacements

Hip 5 Years in Advance Subgraphs		N^{s+}	N^{s-}	Predicted Correctly	Average Influence (%)
NOA →	NSAIDs	**184**	2	184	43.59
Opioids	NSAIDs	**122**	1	122	47.53
Never smoked	Never smoked	**58**	43	58	86.39
Healthy weight	NSAIDs	**54**	0	54	62.49
Obese	Never smoked	**37**	10	37	60.51
Overweight	Never smoked	**21**	26	21	87.49
Overweight	NSAIDs	**38**	5	38	83.74
NSAIDs	HTN	**37**	3	37	54.33
rNSAIDs	rNSAIDs	**39**	0	39	43.89
Obese	Obese	**9**	27	9	90.93

Negative Class

10 Most Frequent Influential Subgraphs for Patients Not Recommended to Have a Hip Replacement

Hip 5 Years in Advance Subgraphs		N^{s+}	N^{s-}	Predicted Correctly	Average Influence (%)
NOA →	NSAIDs	310	**0**	0	36.68
Opioid	NSAIDs	280	**0**	0	34.91
Obese	Never smoked	264	**9**	9	36.59
Never smoked	Never smoked	170	**93**	93	75.90
Ex-smoker	Ex-smoker	180	**10**	10	44.67
Healthy weight	NSAIDs	187	**1**	1	51.43
NOA	Ex-smoker	182	**0**	0	32.19
Obese	Obese	6	**162**	162	97.31
HTN	HTN	0	**119**	119	94.81
Never smoked	Never smoked	109	**6**	6	40.46
Overweight	Overweight				

Figure 7.23: Most influential subgraphs for each class outcome for the hip replacement five years in advance median edge-act method.

7.5 Discussion

The four explainablility methods described in this paper aim to produce a clinician-interpretable justification for each output. These methods may increase decision confidence if the results match clinical expectations. However, the absence of a plausible explanation does not imply an inaccurate model. As these methods are post-hoc, important features may be spuriously correlated with replacement risk [273] or non-causal.

Explainable TG-CNN models are valuable for offering intuitive insights into how clinical code pairs or primary care visits affect predictions, using simple percentages. This accessibility helps medical professionals understand and trust the model. Often 'explainable' methods require advanced technical knowledge, making them inaccessible to users without a ML or data science background [270].

Grad-CAM gave higher weightings to time steps that occurred most recently - this makes sense as more recent things are likely to be more relevant, however this is not hugely valuable to show the influence of clinical codes towards model decision.

The fm-act and Grad-CAM (abs) methods led to graphs with zero sparsity (Table 7.1). This means that these graphs could be difficult to interpret if the patient has a long EHR history. The edge-act median method had the highest sparsity, closely followed by the mean and the maximum version. Therefore, the median edge-act method might be the easiest to interpret for long EHR histories.

The max edge-act method gives the highest MAE between the trained/ original model and the random weighted model, whereas max fm-act gives the lowest MAE. The fm-act method is more prone to EDB than the other methods as shown by its smaller MAE values, and max edge-act is the less likely to have EDB. The Grad-CAM (ReLU) method is less prone to EDB than the Grad-CAM (abs) method.

Figure 7.7 shows a non-Gaussian distribution when comparing maximum activation differences, therefore mean edge-act method was deemed inappropriate. The median edge-act method gives the best sparsity results, whilst the max edge-act method gives the best EDB and slightly higher sensitivity results. The edge-act methods are more sparse than the Grad-CAM (ReLU) model, but this is because it looks at individual edges (clinical code connections) rather than whole visits.

From the results in Table 7.1 it was determined that the median edge-act and Grad-CAM (ReLU) methods provide the best visual explainability for the TG-CNN model. The Grad-CAM (ReLU) model is useful for showing the influence of visits, whilst the edge-act model shows influence of edges. Subgraph frequency analysis was performed on the median edge-act method, as its high sparsity suggests the subgraphs should be smaller and more common amongst individuals. The fm-act method had the most votes for clinical interpretability, however due to lack of sparsity it was believed this method would not be scalable for long EHRs. Clinicians favoured the graphs where the nodes were colour-coded rather than the edges, therefore there might be future scope to adapt the edge influence onto the node colouring.

Looking at the subgraphs produced from the model using the median-edge-act method enabled explainability of the inner workings of the TG-CNN model. Subgraph analysis alongside group classification and accuracy further enables understanding of how subgraphs contribute to probability scores. Some subgraphs appear to always be present in individuals with joint replacement outcomes, whilst others only appear in those without joint replacement outcomes, and similarly some do not appear to be specific.

Interestingly, the hip five years in advance TG-CNN model performed the best and had the lowest number of subgraphs, whilst the knee one year in advance model had the highest number of subgraphs and performed the worst out of the four models. The Pearson correlation coefficient between the total number of subgraphs and the AUROC was -0.90 which indicates a strong negative correlation, meaning that as the number of subgraphs increases, the AUROC tends to decrease.

Given that all models improved with the addition of prescriptions when 200 visits were included and that the majority of the most frequent subgraphs contain prescriptions, it can be inferred that including prescriptions helped the model learn better and find patterns associated to prediction outcomes.

Prescription records, weight, and smoking status were very common in subgraphs in patients with higher hip replacement risks, this could be due to these being the most frequent nodes/ codes occurring in patient EHRs in this dataset. Alternatively, alongside the natural progression of OA and associated co-morbidities other factors may influence surgical decisions. For example smoking status is often recorded for pre-surgical assessments. This is because smoking has been linked to poorer surgical outcomes, including higher risk of wound complications, infection,

aseptic joint loosening, and higher mortality rates [283]. Additionally, other lifestyle factors, such as obesity, can impact patient trajectories and this may be seen in these sub-graphs as these factors can impact surgical eligibility, joint function, and recovery time.

In all datasets there were nearly double the number of subgraphs in the testing sets compared to the training sets, which could explain phenomenon occurring in Chapter 5 where prior to recalibration the curves where inverted. If subgraphs are unseen to a model it may have to infer from subgraphs it has learnt. However, the model with the least numbers of subgraphs performed the best, which could suggest overfitting in the other models.

7.5.1 Limitations

These four methodologies have the following limitations: 1. Due to the nature of these methods, they are not falsifiable without human interpretation. It is unknown whether the model is predicting based on patterns that are reasonable/ align with a clinicians thought process, without clinical assessment. 2. These methods do not consider causality; however the model may help us identify features that influence hip replacement risk that may be currently unknown to clinicians.

To fully determine the association between long patient EHRs and clinician feedback, a correlation analysis would be required to analyse the strength between negative feedback and number of visits in a graph. Due to having only seven clinicians fill out the explainable graph survey, N=7 was too small to perform a correlation analysis. However, future work with larger survey participation would enable correlation analysis.

Long patient histories were too complex for quick visualisation. Clinicians need to see patient trajectories affecting model predictions swiftly. Instead of plotting entire histories, visualising only relevant graphs or subgraphs based on a percentage influence threshold may be more effective.

7.5.2 Future work

There is the potential to add attention layers to give global dependencies across the entire input sequence in combination to the CNN layers which focus on local patterns. Attention may be more suitable for sequential data compared to CNNs which are known for being more useful for hierachial features using spatial data. Similarly to the subgraphs mentioned in this chapter, the

attention maps could be the same shape as filters and subgraphs could be built from these maps. Future work could involve attention mechanisms, allowing the model to focus on specific inputs during the training process. However, these methods may be significantly more computationally expensive.

The TG-CNN model has two 3D CNN branches to deal with time dilation. In this work the first branch which has a stride of 1 was considered, next steps could include analysis and comparison of the second branch.

The feedback on these four suggested methods direct focus to scalability and dimensionality reduction in future iterations of these methodologies, specifically the graph visualisations should be presented more clearly to clinicians, prioritising the most informative regions of the EHR history visually first.

Layer-wise attribution could also be performed to see how much the LSTM layer contributes to the model prediction compared to the CNN layer.

Clinicians can use this tool to assess the five-year risk of a patient needing a hip replacement based on their existing EHR data. For deeper insights into specific model decisions, clinicians can interact with visualisation tools described in this chapter to explore a patient's clinical code history and identify key factors influencing predictions. This can aid in patient care decisions, such as painkiller prescribing, physiotherapy, and exercise recommendations. Additionally, these methods can assist in resource planning by generating lists of patients anticipated to require surgery in five years. Clinicians can show these graphs to their patients, demonstrating model decision making whilst providing motivation for patients to adhere to treatment plans.

7.6 Conclusion

Four methodologies on a temporal graph-based CNN model were developed to improve the explainability of hip replacement risk prediction. The edge-act method provided the best results in terms of graph sparsity, sensitivity, and reduced edge detection bias. Based on subgraph frequency analysis, prescriptions are highly influential to model prediction. Clinicians found these visualisation techniques useful to explain model outputs.

Chapter 8

Conclusions

8.1 Summary of Findings

The main research question for this work was:

Can incorporating elapsed time between EHR events/clinical codes be used, with the TG-CNN model, to improve predictive power and enable clinicians to make informed decisions more effectively, compared to the current state-of-the-art AI approaches?

This question was explored through the following sub-questions, each of which contributed to the outcomes discussed below:

In what ways can graph-based representations and AI improve clinical insights and diagnostic capabilities in MSK research?

Outcome: Graph-based representations were shown to effectively capture the complexities of EHRs, such as irregular time intervals, clinical visits, and diverse codes. This enabled better prediction of hip and knee replacement risks, particularly when combined with CNN architectures.

How are graphs being used on EHRs to predict diagnosis and health outcomes?

Outcome: A systematic exploration of graph representations of EHR within AI models provides competitive performance compared to traditional ML and other existing AI methods.

What is the methodology behind TG-CNN models? Can the TG-CNN model be

used effectively on simple temporal data to predict binary outcomes?

Outcome: The TG-CNN was successfully implemented on temporal MOOC data, achieving state-of-the-art performance metrics for predicting student dropout of online courses.

Can the TG-CNN model be used to predict hip and knee replacement risk at one and five-year intervals, and how does its performance compare to existing models in the literature?

Outcome: The TG-CNN achieved an AUPRC of 0.409 and 0.879 for hip replacement risks at one and five years, respectively, and 0.353 and 0.442 for knee replacement risks. These results compared favourably to existing models, highlighting the utility of graph-based CNN methods for such predictions. Furthermore, addressing imbalanced datasets proved critical for model calibration and reliability. The evaluation on imbalanced unseen test data demonstrated the generalisability of the models across diverse patient populations, reinforcing the applicability of graph-based methods in real-world healthcare settings. While multimodal models incorporating imaging data could potentially enhance prediction, their practicality in primary care is limited by higher costs and complexity. The results emphasize the value of simpler models using routinely collected EHR data for earlier and more cost-effective predictions.

What methods are useful to apply to TG-CNNs to provide explainability for machine-learned predictions to clinicians in a visually understandable way?

Outcome: Interactive graphs and sparse explainability methods were used to extract subgraphs and find areas of EHRs which influenced model decision making, enabling clinicians to visualise global patterns and understand predictions at an individual level. This approach enhanced transparency and practical utility in clinical decision-making.

8.2 Contribution to Knowledge and Gaps Identified

8.2.1 Hip and Knee Replacement Risk in Primary Care

Chapter 6 outlined the current models used in predicting hip and knee replacement risk in advance in primary care. Primary care prediction has much better clinical utility than secondary care prediction as the speed of prognosis is accelerated, such that early detection allows for more conservative and preventative treatments and therapies can be recommended potentially

without secondary care referral. By predicting hip and knee replacement in advance at primary care level, secondary care referral can be triaged to reduce patient travel and time expenditure alongside reducing hospital resource utilisation.

Despite over 20 models being developed for hip and knee replacement risk prediction using secondary care data, only five papers have developed models for this task in primary care data [231, 232, 233, 234, 263]. Two papers covered hip and knee replacement, the other three papers covered knee replacement alone.

Currently the Oxford Knee and Hip Score (OKHS) method is used in clinical practice to screen patients for surgical or non-surgical hip and knee referral [231]. The OKHS method achieved an AUROC of 0.83 and 0.87 for knee and hip referral, respectively. The OKHS method was the best performing model for both hip and knee replacement referrals, that could be found via a systematic search. However, it must be considered that this OKHS model does not give an advance indication of future joint replacement requirement, instead it is used as a clinical tool when a clinician suspects that a joint replacement may be required. The TG-CNN finds early patterns in EHR histories which the model assigns influence to return probabilities up to five years in advance, giving significant time for triaging and intervention treatment plans to begin.

There is very little evidence that any of these models have been fully validated using external datasets and the performance metrics provided, for example AUROC without calibration curves, are poor indicators of robust model performance. To see implementation of these models in clinical practice, researchers must be more transparent with reporting methodology and results, and external validation should be performed on unseen datasets (either in other practices or populations) to ensure generalisability. Clinical trials may then be required to ensure benefit in practice. Regardless, the TG-CNN model outperforms the existing OKHS model by 11.8% and 16.3% in the hip and knee models, respectively. The TRIPOD-AI [186] checklist was followed throughout Chapters 5 and 6 providing transparent methodology.

8.2.2 Modelling using EHRs

Health services face growing challenges due to obesity, diabetes, and ageing populations, creating an urgent need to deliver more with fewer resources. To address this, health systems must be adaptable to the dynamic needs of individuals and populations while optimising resource allocation to maximise health benefits for the greatest number of people [119]. With

the expansion of EHR data and the advancements in modelling techniques, EHR data is being used increasingly for health prediction and modelling tasks to improve clinical decision making, patient outcomes, and planning resource utilisation. Despite these advancements there are still some areas that require improvement and gaps of knowledge to be filled. EHRs enhance clinical research by providing real-world, longitudinal patient data at scale. EHRs are useful for identifying eligible participants, facilitating observational studies, and generating insights into disease progression and treatment outcomes [284]. By integrating research capabilities into clinical workflows, EHRs can bridge the gap between clinical care and research.

Chapter 3 of this thesis presents a systematic review of prediction models using EHRs for forecasting individual patient health outcomes. The review identified only three studies with low RoB, highlighting the importance of researchers adhering to guidelines such as TRIPOD to ensure methodological rigour. Many of these studies inappropriately used clinical codes (for example, including ground truth labels within the model predictors) or predicted outcomes within time frames that were not clinically relevant. Furthermore, none of the models accounted for the irregular time intervals within EHRs, which could provide valuable predictors to enhance model performance. The predicted health outcomes in these studies were limited to mortality, hospital readmission, treatment success, cancer, sepsis, heart failure, heart disease, COPD, URTI, and Alzheimer's disease. This thesis expands on these outcomes by incorporating joint replacement into the list of health outcomes predicted using graph-based representations of individual EHRs.

EHRs have been utilised for various predictive tasks using non-graphical approaches. Among these, ML and deep learning methods are the most commonly employed for disease prediction. Popular techniques include CNNs, RNNs, GNNs, and transformers for analysing EHR data [106, 107, 108]. For risk prediction, generalised models such as LR and Cox regression models remain the most frequently used algorithms [109].

Deep learning and advanced ML techniques are often recommended for their superior performance compared to simpler ML methods, as they can capture more complex relationships [10, 109]. However, the added complexity may hinder their ability to identify and model uncertainties in the data, which poses significant risks in healthcare settings. This issue is particularly critical when the underlying data distribution changes, potentially affecting future predictions [10].

Existing literature reviews examining the use of EHRs in predictive models for health-related outcomes consistently highlight a prevalence of high RoB research. These findings align with the systematic review conducted in this thesis. For instance, one systematic review of AI models using EHR data identified 81 studies [107]. Approximately one-third of these studies exhibited high RoB, with none reporting calibration and most relying on AUROC as a performance metric. Similar patterns were noted in other reviews [106, 109]. Additionally, the scarcity of code availability further undermines trust in these models, compounded by the limited range of performance metrics reported [107].

Another systematic review revealed that out of 81 models predicting risks using EHRs, 34 forecast outcomes within a 90-day timeframe, 32 projected outcomes over a period exceeding one year, and 15 lacked defined endpoints [109].

A gap in the literature appears to exist regarding the incorporation of time-stamp information in AI models using EHR data [107, 108]. Graph-based approaches offer a promising alternative, as they can represent EHRs while preserving structural and temporal relationships through node and edge attributes [144].

Many methods employing EHR data lack transparency and explainability, reducing clinical trust in their predictions [10, 106]. However, attention mechanisms are increasingly being used to address this issue [107, 108]. The graph-based methods described in this thesis enable intuitive transformation of graphs and subgraphs generated by model filters into visualisations, enhancing model explainability and fostering greater trust in their clinical applications.

Ainsworth and Buchan explore the challenges faced by current health systems and propose potential improvements through the use of EHR data [119]. They advocate for reusing existing data to enhance learning health systems. Learning health systems are dynamic models that continuously learn and evolve from updated data to predict health outcomes and manage treatment pathways. Their work introduces conceptualisation of a system aimed at providing real-time, personalised care recommendations to support clinical decision-making by accounting for individual patient differences, thereby optimising outcomes. However, one major obstacle is the fragmentation of data, often siloed regionally, by domain (e.g., education, healthcare, employment), or by institution/hospital. The lack of interoperability across healthcare systems, hampers the integration between datasets. Initiatives like Connected Bradford, the Combined Intelligence for Population Health Action (CIPHA) system, and the Greater Manchester (GM)

Care Record are gaining traction by integrating data sources to provide a comprehensive view of an individual's lifestyle and health [285]. Despite such advancements, significant challenges remain, including addressing privacy concerns surrounding the use of sensitive data and the need for technological advancements to facilitate widespread implementation. Using EHRs in population health management can help address challenges that continue to rise. For example, risk stratification, preventative care, disease management, and determining health and social disparities. For this to happen, models using EHRs need to be robust and generalisable which provide decision making and planning far enough in advance to be clinically relevant.

8.3 Implications of the Research

This work introduces a completely new type of model for predicting outcomes from temporal sequences, enabling the incorporation of irregular time intervals and non-linear event sequences for greater representation and learning. The models developed in this work demonstrate the potential for forecasting the likelihood of hip or knee replacement in one or five years. By identifying individuals at risk, these tools could support proactive clinical interventions aimed at preventing or delaying joint replacement through targeted management strategies. Their implementation could enhance operational planning by anticipating future surgical demands and allocating resources more effectively. Integrating such predictive models into clinical workflows represents an opportunity to improve patient outcomes and optimise healthcare delivery.

The integration of predictive models, such as those developed in this thesis, into learning health systems could significantly enhance service planning at the council and Integrated Care Board (ICB) levels. By using real-time data and iterative model updates, these systems can identify patients at risk of hip or knee replacements well in advance, allowing for more strategic allocation of healthcare resources. For example, areas with high predicted replacement rates could proactively plan for increased surgical capacity, rehabilitation services, and physiotherapy programs. These integrated models will need to be updated to ensure things such as temporal drift, when treatment or healthcare practices change altering model predictive performance, do not occur. Reduction in model performance, significant changes to incidence rates or elapsed time could trigger the need to re-train the model. Ideally this model should be less computationally expensive to ensure more frequent model updates (retraining or recalibration) so that prediction accuracy does not decline.

Predictive insights could inform care home utilisation, helping councils and ICBs anticipate and accommodate patients who may require short-term residential care following surgery. By embedding risk prediction models into learning health systems, stakeholders can move from reactive to proactive care planning, helping cost efficiency, improving patient outcomes and reducing strain on healthcare systems [119]. Such integration exemplifies the transformative potential of predictive analytics in addressing the broader challenges of ageing populations and rising demand for orthopaedic care.

The methodology described in this thesis is adaptable to new clinical domains using temporal data. The intuitive 3-tensor construction and representation is versatile for explainable methods which could be further experimented with and explored for user experience.

8.4 Limitations of the Research

In this section the main limitations are covered, further research limitations are provided in the previous chapters.

8.4.1 Literature Bias

This research systematically investigated graph-based and ML methods applied to EHRs for health outcome prediction. However, only one relevant low-RoB paper was identified, which shifted the chapter's focus to methodological bias. Although these highlighted areas require bias reduction, it promotes the need for careful consideration of assumptions to ensure clinical applicability of such methods.

The use of the PROBAST tool, while a gold standard for assessing RoB in traditional predictive modelling, may inadequately address the unique challenges of ML, particularly deep learning methods. New PROBAST-AI guidelines, aim to better appraise ML methods, but it is anticipated that the findings of this study, where most papers had high RoB in the analysis domain, will remain consistent under these forthcoming frameworks.

8.4.2 Evolving Research Landscape

One of the key limitations of this thesis is the rapidly evolving nature of the field, where new studies and models using graph-based approaches in EHRs are frequently emerging. Figure 3.2 from Chapter 3 show the growth of graphs in healthcare prediction research since the

2000's. As researchers continue to explore novel methodologies, the techniques and findings presented here may quickly become outdated or require adaptation in response to more advanced frameworks. This constant influx of new research presents both a challenge and an opportunity, it underscores the need for continuous refinement and highlights the dynamic progress in EHR-based graph modelling. Future studies may build upon, refine, or even challenge the assumptions and methodologies used in this thesis.

For example, a paper introducing temporal graph transformers has been introduced which enables the inclusion of temporal EHR data [286]. This paper captures both visit-level and event-level interactions, enabling more complex healthcare system interactions to be captured. This work also focuses on node/edge-level prediction, which is not highly useful for pre-planning or preventative care. The authors also do not include the outcomes to be predicted which makes it difficult to statistically evaluate. Whilst this paper differs greatly to the work carried out in this thesis, it demonstrates that research in this field will continue to grow and that continuous revisiting of the literature is required if the work in this thesis is to be furthered.

8.4.3 Challenges of Using Clinical Codes

The EHR dataset used may introduce inaccuracies due to inconsistencies in coding and event recording dates, potentially affecting the reliability of predictive features. Similarly, left- and right-censoring issues may misclassify hip replacement statuses.

Knee OA diagnoses are often delayed in EHRs, with narrative diagnoses preceding codified records by years. This highlights the need to investigate the incorporation of narrative data into predictive models to enhance accuracy and timeliness.

Joint replacement dates may not have been coded to date, which means that waiting times and insights into optimal surgery times may be skewed.

CTV3 codes are used in model training, however the transition from the CTV3 to SNOMED-CT coding system does not affect the methodology described in this thesis, which is adaptable to different coding systems.

8.4.4 Dataset Challenges

The dataset lacked critical features, such as ethnicity and laterality (e.g., left or right limb), which could improve model granularity and explainability. Socioeconomic status was repre-

sented by IMD quintiles derived from postcodes, potentially misrepresenting individual deprivation levels. Additionally, medical reasons for replacements (including emergency procedure records, such as after fractures) or their elective nature were not available, limiting insights into underlying health conditions. Furthermore, the dataset used was smaller and less comprehensive than other datasets, such as CPRD.

This dataset was not provided with information when patients were lost to follow-up or changed practice therefore right-censoring could mean that a patient would have had a hip replacement, but the label is inaccurately given as not needing a replacement.

Using ResearchOne data leads to further limitations. ResearchOne data relies on anonymised EHRs, which may not include all relevant patient histories due to patient's opting out of sharing their data leading to data incompleteness [70]. As patient's opt out the sample size decreases, potentially leading to biases in the model [287]. Opt-outs can disproportionately affect certain demographics, which may cause the trained model to lack representative patterns, reducing model generalisability [70]. The temporal graph-based models used in this project rely on tracking patient primary care visits over time, but the combination of missing records or opt-outs occurring could lead make it more difficult to the model to converge with more gaps in patient trajectories [288].

ResearchOne is a primary care database in the UK that collects data from general practices using the SystmOne clinical system. While SystmOne is widely used across England, its coverage is geographically uneven, meaning ResearchOne may not provide a fully representative sample of the UK population. Certain regions where SystmOne is more prevalent may be overrepresented, while others where alternative systems are dominant may be underrepresented. This potential skew in representation could introduce biases in research findings, this may reduce generalisability due to regional variations in healthcare outcomes and demographics [289].

CPRD collects data from a broader set of practices using various clinical systems, making it a more representative source of primary care data in the UK [290]. The Health Improvement Network (THIN), based on the Vision system, has a more limited geographical spread, while QResearch, which relies on EMIS practices, has strong representation in London, the South, and the West of England [291]. The choice of database can significantly affect the representativeness of research findings, so we must consider these differences when selecting data sources.

Despite out-of-sample testing of the recalibrated model showing good performance, external validation using unseen datasets was not performed, limiting the generalisability of the models. Best practices would involve using separate datasets or splitting by time or health centres, for multi-centre studies.

8.4.5 Explainability Limitations

Explainability methods used in this study are not falsifiable without human interpretation and do not consider causality, posing challenges in aligning model predictions with clinical reasoning. Careful wording to clinicians is required to ensure that they understand visit or clinical code influence are based on prediction from the model than meaningful or causal pathways.

The small sample size (N=7) of clinicians in the graph visualisation survey prevented correlation analysis of explainability feedback, requiring larger participation in future studies.

Long patient histories were too complex for effective visualisation. Simplified visualisations, such as subgraphs filtered by influence thresholds or initial focus on more influential areas, may offer more practical insights for clinicians.

Whilst the research undertaken for this thesis showed improvements to predicting hip and knee replacement risk with tools which could be clinical beneficial for assisting with decision making, it must be acknowledged that this work has a long way to go. Clinicians in the UK working for the NHS currently have limited equipment and resources. Before healthcare joins the AI revolution, serious thought is needed into resourcing and updating existing basic equipment (ECG machines, sphygmomanometers).

8.5 Recommendations for Future Research

To demonstrate the real-world impact of this model, future work should evaluate its clinical utility through metrics such as patient quality of life, clinician workflow improvements, and healthcare resource utilisation. Decision curve analysis could provide insights into the balance between benefit and harm based on clinical risk thresholds, addressing questions like: “What are the consequences of a model incorrectly predicting an outcome? Would such errors significantly impact patient health, either by unnecessary interventions or missed opportunities for necessary care?” By weighting these factors, decision curve analysis can identify thresholds that minimise

harm and maximise benefit, accounting for variations in clinician and patient perceptions of life impact. Decision curve analysis combined with a meta-analysis of interventions may also reveal cost-effective strategies for reducing healthcare expenditures.

Integrating a Patient and Public Involvement and Engagement (PPIE) group could help refine risk grouping criteria, ensuring alignment with patient needs. The TG-CNN model could be adapted for triaging patients, integrating into primary care to provide real-time risk scores for hip or knee replacements based on joint pain codes in the EHR. These scores could inform follow-up care strategies and resource planning.

Currently, graph-based methods in healthcare are underutilised, limited to seven prediction outcomes (mortality, readmission, treatment success, sepsis, cardiovascular disease, Alzheimer's, and now joint replacement). Future work should explore broader health applications, such as predicting resource utility or cancer recurrence.

External validation of the model is critical to assess its generalisability. This involves testing the model on datasets from different regions or health systems to ensure robust performance.

Future work should enhance the explainability of the TG-CNN model to assist clinically informed decision making about patient care. Visualisations should be simplified for clinical use, presenting only the most relevant aspects of patient histories to improve clarity and usability. Visual tools could enhance transparency by illustrating the reasoning behind risk scores, empowering patients to understand their health risks and motivating adherence to treatment plans. It may be worth exploring attention mechanisms to improve interpretability. Attention layers could complement the existing CNN architecture by capturing global dependencies in sequential data, making predictions more understandable. Subgraphs derived from attention maps could prioritise the most impactful regions of a patient's EHR for visualisation. However, it must be considered that attention mechanisms may increase computational costs. Additionally, layer-wise attribution could help determine the contributions of the CNN and LSTM layers to model predictions. For these models to be used in practice impact assessments would need to be carried out which involve randomised controlled trials between regions. Some regions will have the models in place to trial and others will not use a model and be the control group.

Edge-case analysis may also be beneficial, looking at the graphs from individuals who have been incorrectly predicted due to rare health trajectories. These edge-case may be useful for case

study analysis by clinically trained professionals, to determine if the trajectory is rare both to the predictive model and a clinician. These cases can help refine models, as they reveal weaknesses in generalisation or biases in training data. Continuous improvement could involve logging edge cases, analysing their impact, and retraining models with adjusted parameters or additional data to enhance robustness.

Future iterations of graph-based visualizations should focus on scalability and dimensionality reduction to address the complexity of long patient histories. Graphs should emphasize the most informative regions of the EHR, presenting concise and actionable insights to clinicians.

Further investigation into the dual-branch TG-CNN architecture could explore whether the second branch, designed to handle time dilation with a different stride, improves visualization and interpretability. This could provide deeper insights into how the model captures temporal patterns in the data.

By addressing these areas, future research can enhance the model's generalisability, clinical relevance, and usability, bridging the gap between predictive analytics and real-world healthcare applications.

8.6 Final Remarks

This thesis shows the potential of integrating advanced ML and graph-based methods with EHRs to predict health outcomes. By critically analysing current methodologies and highlighting key challenges, it sets a foundation for addressing biases and enhancing the reliability and interpretability of predictive models. These contributions not only push the boundaries of data-driven healthcare but also emphasise the ethical and clinical responsibility of ensuring these models are both equitable and actionable.

While the limitations and challenges identified highlight the complexity of translating predictive analytics into clinical practice, they also present opportunities for innovation. The importance of external validation, patient-centric model development, and explainable AI methods cannot be overstated. By focusing on these areas, future research can deliver solutions that are not just technically sound but also tailored to real-world healthcare needs.

This work envisions a healthcare ecosystem where predictive models empower clinicians and patients alike, enabling earlier interventions, personalised care plans, and improved outcomes.

The pathway from research to implementation is not without hurdles, but with continued focus on methodological rigour, clinical relevance, and stakeholder collaboration, the promise of predictive analytics in reshaping healthcare can become a reality. This thesis is a step forward in that journey, laying the groundwork for models that are not only innovative but also trustworthy, transparent, and transformative.

References

[1] Davenport, T. and Kalakota, R. "The Potential for Artificial Intelligence in Healthcare". In: *SSRN Electronic Journal* 6.2 (2020), pp. 94–98. DOI: 10.2139/ssrn.3525037.

[2] Secretary of State Science Innovation and Technology. *AI Opportunities Action Plan.* Tech. rep. Department for Science, Innovation & Technology, 2025. URL: https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan.

[3] Sun, C., Hong, S., Song, M., and Li, H. "A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data". In: *arXiv preprint* (2020). URL: https://arxiv.org/abs/2010.12493.

[4] Whitty, C. *Chief Medical Officer's annual report 2020: health trends and variation in England.* Tech. rep. NHS, 2020, pp. 1–120. URL: https://www.gov.uk/government/publications/chief-medical-officers-annual-report-2020-health-trends-and-variation-in-england.

[5] Conaghan, P. G., Kloppenburg, M., Schett, G., and Bijlsma, J. W. J. "Osteoarthritis research priorities : a report from a EULAR ad hoc expert committee". In: (2014), pp. 1442–1445. DOI: 10.1136/annrheumdis-2013-204660.

[6] Cook, M. J., Lunt, M., Ashcroft, D. M., Board, T., and O'Neill, T. W. "The Impact of Frailty and Deprivation on the Likelihood of Receiving Primary Total Hip and Knee Arthroplasty among People with Hip and Knee Osteoarthritis". In: *Journal of Frailty and Aging* 12.4 (2023), pp. 298–304. ISSN: 22734309. DOI: 10.14283/jfa.2023.36.

[7] National Joint Registry. *National Joint Registry 21st Annual Report 2024: Hip Replacements*. Tech. rep. National Joint Registry, 2024. URL: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2021st%20Annual%20Report%202024_Hips.pdf.

[8] Svege, I., Nordsletten, L., Fernandes, L., and Risberg, M. A. "Exercise therapy may postpone total hip replacement surgery in patients with hip osteoarthritis: A long-term follow-up of a randomised trial". In: *Annals of the Rheumatic Diseases* 74.1 (2015), pp. 164–169. ISSN: 14682060. DOI: 10.1136/annrheumdis-2013-203628.

[9] Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., and Thiesson, B. "Explainable artificial intelligence model to predict acute critical illness from electronic health records". In: *Nature Communications* 11.1 (July 2020), p. 3852. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17431-x. URL: https://www.nature.com/articles/s41467-020-17431-x.

[10] Xiao, C., Choi, E., and Sun, J. "Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review". In: *Journal of the American Medical Informatics Association* 25[1] C. X.10 (2018), pp. 1419–1428. ISSN: 1527974X. DOI: 10.1093/jamia/ocy068.

[11] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da.* Tech. rep. European Parliament and Council, 2016, pp. 1–88. URL: https://gdpr-info.eu/.

[12] Matheny, M., Israni, S. T., Ahmed, M., and Whicher, D. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Ed. by Matheny, M., Israni, S. T., Ahmed, M., and Whicher, D. Washington, D.C.: National Academies Press, Aug. 2019. ISBN: 978-0-309-70513-4. DOI: 10.17226/27111. URL: https://www.nap.edu/catalog/27111.

[13] Smith, C., Hewison, J., West, R. M., Kingsbury, S. R., and Conaghan, P. G. "Understanding patterns of care for musculoskeletal patients using routinely collected National

Health Service data from general practices in England". In: *Health Informatics Journal* 26.4 (Dec. 2020), pp. 2470–2484. ISSN: 1460-4582. DOI: 10.1177/1460458220907431. URL: http://journals.sagepub.com/doi/10.1177/1460458220907431.

[14] Yu, D., Jordan, K. P., Snell, K. I., Riley, R. D., Bedson, J., Edwards, J. J., Mallen, C. D., Tan, V., Ukachukwu, V., Prieto-Alhambra, D., Walker, C., and Peat, G. "Development and validation of prediction models to estimate risk of primary total hip and knee replacements using data from the UK: Two prospective open cohorts using the UK Clinical Practice Research Datalink". In: *Annals of the Rheumatic Diseases* 78.1 (2019), pp. 91–99. ISSN: 14682060. DOI: 10.1136/annrheumdis-2018-213894.

[15] Li, H., Chan, L., Chan, P., and Wen, C. "An interpretable knee replacement risk assessment system for osteoarthritis patients". In: *Osteoarthritis and Cartilage Open* 6.2 (June 2024), p. 100440. ISSN: 26659131. DOI: 10.1016/j.ocarto.2024.100440. URL: https://linkinghub.elsevier.com/retrieve/pii/S2665913124000074.

[16] Versus Arthritis. *The State of Musculoskeletal Health 2023 Arthritis and Other Musculoskeletal Conditions in Numbers*. Tech. rep. Versus Arthritis, 2023, pp. 25–28. URL: https://www.versusarthritis.org/media/25650/versus-arthritis-state-msk-musculoskeletal-health-2023-accessible.docx#:~:text=Every%20year%201%20IN%205,to%20help%20manage%20their%20symptoms..

[17] Parsons, S. and Symmons, D. P. "The burden of musculoskeletal conditions". In: *Medicine* 38.3 (Mar. 2010), pp. 126–128. ISSN: 13573039. DOI: 10.1016/j.mpmed.2009.11.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1357303909003442.

[18] Keavy, R. "The prevalence of musculoskeletal presentations in general practice: an epidemiological study". In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 70 (2020), pp. 1–7. ISSN: 14785242. DOI: 10.3399/bjgp20X711497.

[19] Cieza, A., Causey, K., Kamenov, K., Hanson, S. W., Chatterji, S., and Vos, T. "Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019". In: *The Lancet* 396.10267 (Dec. 2020), pp. 2006–2017. ISSN: 01406736. DOI: 10.1016/S0140-

6736(20)32340-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673620323400.

[20] Briggs, A. M., Woolf, A. D., Dreinhöfer, K., Homb, N., Hoy, D. G., Kopansky-Giles, D., Åkesson, K., and March, L. "Reducing the global burden of musculoskeletal conditions". In: *Bulletin of the World Health Organization* 96.5 (May 2018), pp. 366–368. ISSN: 0042-9686. DOI: 10.2471/BLT.17.204891. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5985424/pdf/BLT.17.204891.pdf/.

[21] NICE. *Osteoarthritis: care and management*. Tech. rep. Nice Institute for Health and Care Excellence (NICE), 2022. URL: https://www.nice.org.uk/guidance/cg177.

[22] Herzog, M. M., Driban, J. B., Cattano, N. M., Kenneth, L., Tourville, T. W., and Marshall, S. W. "the Osteoarthritis Initiative". In: 44.8 (2018), pp. 1265–1270. DOI: 10.1002/jmri.25892.Tool.

[23] Dell'Isola, A., Allan, R., Smith, S. L., Marreiros, S. S. P., and Steultjens, M. "Identification of clinical phenotypes in knee osteoarthritis: a systematic review of the literature". In: *BMC Musculoskeletal Disorders* 17.1 (Dec. 2016), p. 425. ISSN: 1471-2474. DOI: 10.1186/s12891-016-1286-2. URL: http://bmcmusculoskeletdisord.biomedcentral.com/articles/10.1186/s12891-016-1286-2.

[24] Zhang, Y. and Jordan, J. M. "Epidemiology of osteoarthritis." In: *Clinics in geriatric medicine* 26.3 (Aug. 2010), pp. 355–69. ISSN: 1879-8853. DOI: 10.1016/j.cger.2010.03.001. URL: http://www.ncbi.nlm.nih.gov/pubmed/20699159%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2920533.

[25] Versus Arthritis. *The Musculoskeletal Calculator*. Tech. rep. Versus Arthritis, 2023. URL: https://www.versusarthritis.org/policy/resources-for-policy-makers/musculoskeletal-calculator/.

[26] Chamberlain, R. "Hip Pain in Adults". In: *American family physician* 103.2 (2021), pp. 81–89. ISSN: 15320650.

[27] Conaghan, P. G., D'Agostino, M. A., Le Bars, M., Baron, G., Schmidely, N., Wakefield, R., Ravaud, P., Grassi, W., Martin-Mola, E., So, A., Backhaus, M., Malaise, M., Emery,

P., and Dougados, M. "Clinical and ultrasonographic predictors of joint replacement for knee osteoarthritis: results from a large, 3-year, prospective EULAR study". In: *Annals of the Rheumatic Diseases* 69.4 (Apr. 2010), pp. 644–647. ISSN: 0003-4967. DOI: 10.1136/ard.2008.099564. URL: https://ard.bmj.com/lookup/doi/10.1136/ard.2008.099564.

[28] Yu, D., Missen, M., Jordan, K. P., Edwards, J. J., Bailey, J., Wilkie, R., Fitzpatrick, J., Ali, N., Niblett, P., and Peat, G. "Trends in the Annual Consultation Incidence and Prevalence of Low Back Pain and Osteoarthritis in England from 2000 to 2019: Comparative Estimates from Two Clinical Practice Databases". In: *Clinical Epidemiology* 14.February (2022), pp. 179–189. ISSN: 11791349. DOI: 10.2147/CLEP.S337323.

[29] Losina, E., Weinstein, A. M., Reichmann, W. M., Burbine, S. A., Solomon, D. H., Daigle, M. E., Rome, B. N., Chen, S. P., Hunter, D. J., Suter, L. G., Jordan, J. M., and Katz, J. N. "Lifetime Risk and Age at Diagnosis of Symptomatic Knee Osteoarthritis in the US". In: *Arthritis Care & Research* 65.5 (May 2013), pp. 703–711. ISSN: 2151-464X. DOI: 10.1002/acr.21898. URL: https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.21898.

[30] Kolasinski, S. L., Neogi, T., Hochberg, M. C., Oatis, C., Guyatt, G., Block, J., Callahan, L., Copenhaver, C., Dodge, C., Felson, D., Gellar, K., Harvey, W. F., Hawker, G., Herzig, E., Kwoh, C. K., Nelson, A. E., Samuels, J., Scanzello, C., White, D., Wise, B., Altman, R. D., DiRenzo, D., Fontanarosa, J., Giradi, G., Ishimori, M., Misra, D., Shah, A. A., Shmagel, A. K., Thoma, L. M., Turgunbaev, M., Turner, A. S., and Reston, J. "2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee." In: *Arthritis care & research* 72.2 (Feb. 2020), pp. 149–162. ISSN: 2151-4658. DOI: 10.1002/acr.24131. URL: http://www.ncbi.nlm.nih.gov/pubmed/31908149%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC11488261.

[31] Peat, G., Greig, J., Wood, L., Wilkie, R., Thomas, E., and Croft, P. *Diagnostic discordance: We cannot agree when to call knee pain 'osteoarthritis'*. 2005. DOI: 10.1093/fampra/cmh702.

[32] Ferguson, R. J., Prieto-Alhambra, D., Walker, C., Yu, D., Valderas, J. M., Judge, A., Griffiths, J., Jordan, K. P., Peat, G., Glyn-Jones, S., and Silman, A. J. "Validation of hip osteoarthritis diagnosis recording in the UK Clinical Practice Research Datalink". In: *Pharmacoepidemiology and Drug Safety* 28.2 (2019), pp. 187–193. ISSN: 10991557. DOI: 10.1002/pds.4673.

[33] Crawford, R. W. and Murray, D. W. "Total hip replacement: indications for surgery and risk factors for failure". In: *Annals of the Rheumatic Diseases* 56.8 (Aug. 1997), pp. 455–457. ISSN: 0003-4967. DOI: 10.1136/ard.56.8.455. URL: https://ard.bmj.com/lookup/doi/10.1136/ard.56.8.455.

[34] Kellgren, J. H. and Lawrence, J. S. "Radiological assessment of osteo-arthrosis." In: *Annals of the rheumatic diseases* 16.4 (Dec. 1957), pp. 494–502. ISSN: 0003-4967. DOI: 10.1136/ard.16.4.494. URL: http://www.ncbi.nlm.nih.gov/pubmed/13498604%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1006995.

[35] Schiphof, D., Boers, M., and Bierma-Zeinstra, S. M. A. "Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis". In: *Annals of the Rheumatic Diseases* 67.7 (July 2008), pp. 1034–1036. ISSN: 0003-4967. DOI: 10.1136/ard.2007.079020. URL: https://ard.bmj.com/lookup/doi/10.1136/ard.2007.079020.

[36] Stubbs, B., Aluko, Y., Myint, P. K., and Smith, T. O. "Prevalence of depressive symptoms and anxiety in osteoarthritis: a systematic review and meta-analysis". In: *Age and Ageing* 45.2 (Mar. 2016), pp. 228–235. ISSN: 0002-0729. DOI: 10.1093/ageing/afw001. URL: https://academic.oup.com/ageing/article-lookup/doi/10.1093/ageing/afw001.

[37] Murphy, L. and Helmick, C. G. "The Impact of Osteoarthritis in the United States". In: *Orthopaedic Nursing* 31.2 (Mar. 2012), pp. 85–91. ISSN: 0744-6020. DOI: 10.1097/NOR.0b013e31824fcd42. URL: https://journals.lww.com/00006416-201203000-00006.

[38] The NJR Editorial Board. *National Joint Registry 12th Annual Report*. Tech. rep. National Joint Registry, 2015, pp. 33–35. URL: https://reports.njrcentre.org.uk/Portals/3/PDFdownloads/NJR%2012th%20Annual%20Report%202015.pdf.

[39] Hobbs, F. D., Bankhead, C., Mukhtar, T., Stevens, S., Perera-Salazar, R., Holt, T., and Salisbury, C. "Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007–14". In: *The Lancet* 387.10035 (2016), pp. 2323–2330. ISSN: 1474547X. DOI: 10.1016/S0140-6736(16)00620-6.

[40] Briggs, T. *A national review of adult elective orthopaedic services in England. Getting it right first time*. Tech. rep. 2015. URL: https://gettingitrightfirsttime.co.uk/girft-reports/.

[41] Salaffi, F., Farah, S., and Di Carlo, M. "Frailty syndrome in rheumatoid arthritis and symptomatic osteoarthritis: an emerging concept in rheumatology." In: *Acta bio-medica : Atenei Parmensis* 91.2 (May 2020), pp. 274–296. ISSN: 2531-6745. DOI: 10.23750/abm.v91i2.9094. URL: http://www.ncbi.nlm.nih.gov/pubmed/32420963%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7569610.

[42] NHS England. *CCG Programme Budgeting Benchmarking Tool 2013/14*. Tech. rep. NHS England, 2015.

[43] Office for National Statistics. *Sickness absence in the UK labour market 2021*. Tech. rep. Office for National Statistics, 2021. URL: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/labourproductivity/articles/sicknessabsenceinthelabourmarket/2021.

[44] Rajah, N., Webb, E. J. D., Hulme, C., Kingsbury, S. R., West, R., and Martin, A. "How does arthritis affect employment? Longitudinal evidence on 18,000 British adults with arthritis compared to matched controls." In: *Social science & medicine (1982)* 321 (Mar. 2023), p. 115606. ISSN: 1873-5347. DOI: 10.1016/j.socscimed.2022.115606. URL: http://www.ncbi.nlm.nih.gov/pubmed/36732169.

[45] Katz, J. "Total joint replacement in osteoarthritis". In: *Best Practice & Research Clinical Rheumatology* 20.1 (Feb. 2006), pp. 145–153. ISSN: 15216942. DOI: 10.1016/j.berh.2005.09.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S152169420500104X.

[46] Davis, A., Perruccio, A., Ibrahim, S., Hogg-Johnson, S., Wong, R., Streiner, D., Beaton, D., Côté, P., Gignac, M., Flannery, J., Schemitsch, E., Mahomed, N., and Badley, E. "The trajectory of recovery and the inter-relationships of symptoms, activity and participation in the first year following total hip and knee replacement". In: *Osteoarthritis and Cartilage* 19.12 (Dec. 2011), pp. 1413–1421. ISSN: 10634584. DOI: 10.1016/j.joca.2011.08.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1063458411002408.

[47] Burn, E., Murray, D., Hawker, G., Pinedo-Villanueva, R., and Prieto-Alhambra, D. "Lifetime risk of knee and hip replacement following a GP diagnosis of osteoarthritis: a real-world cohort study". In: *Osteoarthritis and Cartilage* 27.11 (Nov. 2019), pp. 1627–1635. ISSN: 10634584. DOI: 10.1016/j.joca.2019.06.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S106345841931088X.

[48] Appleyard, T., Ashworth, J., Bedson, J., Yu, D., and Peat, G. "Trends in gabapentinoid prescribing in patients with osteoarthritis: a United Kingdom national cohort study in primary care". In: *Osteoarthritis and Cartilage* 27.10 (2019), pp. 1437–1444. ISSN: 15229653. DOI: 10.1016/j.joca.2019.06.008.

[49] Varacallo, M., Luo, T. D., Mabrouk, A., and Johanson, N. A. *Total Knee Arthroplasty Techniques*. 2024. URL: http://www.ncbi.nlm.nih.gov/pubmed/30224001%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8414145.

[50] Jameson, S. S., Mason, J., Baker, P. N., Gregg, P. J., Deehan, D. J., and Reed, M. R. "Implant Optimisation for Primary Hip Replacement in Patients over 60 Years with Osteoarthritis: A Cohort Study of Clinical Outcomes and Implant Costs Using Data from England and Wales". In: *PLOS ONE* 10.11 (Nov. 2015). Ed. by Pérez, M. A., e0140309. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0140309. URL: https://dx.plos.org/10.1371/journal.pone.0140309.

[51] Matharu, G., Culliford, D., Blom, A., and Judge, A. "Projections for primary hip and knee replacement surgery up to the year 2060: an analysis based on data from The National Joint Registry for England, Wales, Northern Ireland and the Isle of Man". In: *The Annals of The Royal College of Surgeons of England* 104.6 (June 2022), pp. 443–

448. ISSN: 0035-8843. DOI: 10.1308/rcsann.2021.0206. URL: https://publishing.rcseng.ac.uk/doi/10.1308/rcsann.2021.0206.

[52] National Joint Registry. *11th Annual Report 2014 National Joint Registry for England, Wales and Northern Ireland*. Tech. rep. 2014. URL: https://reports.njrcentre.org.uk/Portals/1/PDFdownloads/NJR%2011th%20Annual%20Report%202014.pdf.

[53] National Joint Registry. *National Joint Registry 21st Annual Report 2024: Knee Replacements*. Tech. rep. 2024. URL: https://reports.njrcentre.org.uk/Portals/0/PDFdownloads/NJR%2021st%20Annual%20Report%202024_Knees.pdf.

[54] Abdelaal, M. S., Restrepo, C., and Sharkey, P. F. "Global Perspectives on Arthroplasty of Hip and Knee Joints". In: *Orthopedic Clinics of North America* 51.2 (Apr. 2020), pp. 169–176. ISSN: 00305898. DOI: 10.1016/j.ocl.2019.11.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0030589819301725.

[55] Kerrigan, D. C., Todd, M. K., and Della Croce, U. "Gender differences in joint biomechanics during walking: normative study in young adults." In: *American journal of physical medicine & rehabilitation* 77.1 (1998), pp. 2–7. ISSN: 0894-9115. DOI: 10.1097/00002060-199801000-00002. URL: http://www.ncbi.nlm.nih.gov/pubmed/9482373.

[56] Hawker, G. A., Wright, J. G., Coyte, P. C., Williams, J. I., Harvey, B., Glazier, R., and Badley, E. M. "Differences between Men and Women in the Rate of Use of Hip and Knee Arthroplasty". In: *New England Journal of Medicine* 342.14 (Apr. 2000), pp. 1016–1022. ISSN: 0028-4793. DOI: 10.1056/NEJM200004063421405. URL: http://www.nejm.org/doi/abs/10.1056/NEJM200004063421405.

[57] Goodman, S. M., Ramsden-Stein, D. N., Huang, W.-T., Zhu, R., Figgie, M. P., Alexiades, M. M., and Mandl, L. A. "Patients with Rheumatoid Arthritis Are More Likely to Have Pain and Poor Function After Total Hip Replacements than Patients with Osteoarthritis". In: *The Journal of Rheumatology* 41.9 (Sept. 2014), pp. 1774–1780. ISSN: 0315-162X. DOI: 10.3899/jrheum.140011. URL: http://www.jrheum.org/lookup/doi/10.3899/jrheum.140011.

[58] Johnsen, M. B., Hellevik, A. I., Baste, V., Furnes, O., Langhammer, A., Flugsrud, G., Nordsletten, L., Zwart, J. A., and Storheim, K. "Leisure time physical activity and the risk of hip or knee replacement due to primary osteoarthritis: a population based cohort study (The HUNT Study)". In: *BMC Musculoskeletal Disorders* 17.1 (Dec. 2016), p. 86. ISSN: 1471-2474. DOI: 10.1186/s12891-016-0937-7. URL: http://www.biomedcentral.com/1471-2474/17/86.

[59] Yang, Y., Komisar, V., Shishov, N., Lo, B., Korall, A. M., Feldman, F., and Robinovitch, S. N. "The Effect of Fall Biomechanics on Risk for Hip Fracture in Older Adults: A Cohort Study of Video-Captured Falls in Long-Term Care". In: *Journal of Bone and Mineral Research* 35.10 (Dec. 2020), pp. 1914–1922. ISSN: 0884-0431. DOI: 10.1002/jbmr.4048. URL: https://academic.oup.com/jbmr/article/35/10/1914/7516811.

[60] Evans, J. T., Evans, J. P., Walker, R. W., Blom, A. W., Whitehouse, M. R., and Sayers, A. "How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up". In: *The Lancet* 393.10172 (Feb. 2019), pp. 647–654. ISSN: 01406736. DOI: 10.1016/S0140-6736(18)31665-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673618316659.

[61] Evans, J. T., Walker, R. W., Evans, J. P., Blom, A. W., Sayers, A., and Whitehouse, M. R. "How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up". In: *The Lancet* 393.10172 (Feb. 2019), pp. 655–663. ISSN: 01406736. DOI: 10.1016/S0140-6736(18)32531-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673618325315.

[62] Li, Y., Qian, B., Zhang, X., and Liu, H. "Knowledge guided diagnosis prediction via graph spatial-temporal network". In: *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020*. 2020, pp. 19–27. ISBN: 9781611976236. DOI: 10.1137/1.9781611976236.3.

[63] Wu, R., Ma, Y., Yang, Y., Li, M., Zheng, Q., and Fu, G. "A clinical model for predicting knee replacement in early-stage knee osteoarthritis: data from osteoarthritis initiative".

In: *Clinical Rheumatology* 41.4 (Apr. 2022), pp. 1199–1210. ISSN: 0770-3198. DOI: 10.1007/s10067-021-05986-z. URL: https://link.springer.com/10.1007/s10067-021-05986-z.

[64] NHS. *The NHS Long Term Plan*. Tech. rep. NHS, 2019, pp. 1–136. URL: https://www.longtermplan.nhs.uk/publication/nhs-long-term-plan/.

[65] Wells, B. J., Nowacki, A. S., Chagin, K., and Kattan, M. W. "Strategies for Handling Missing Data in Electronic Health Record Derived Data". In: *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 1.3 (2013), p. 7. DOI: 10.13063/2327-9214.1035.

[66] Zhang, Y., Yang, X., Ivy, J., and Chi, M. "Attain: Attention-based time-aware LSTM networks for disease progression modeling". In: *IJCAI International Joint Conference on Artificial Intelligence* 2019-Augus (2019), pp. 4369–4375. ISSN: 10450823. DOI: 10.24963/ijcai.2019/607.

[67] Chiew, C. J., Liu, N., Wong, T. H., Sim, Y. E., and Abdullah, H. R. "Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission". In: *Annals of surgery* 272.6 (2020), pp. 1133–1139. ISSN: 15281140. DOI: 10.1097/SLA.0000000000003297.

[68] Zong, N., Ngo, V., Stone, D. J., Wen, A., Zhao, Y., Yu, Y., Liu, S., Huang, M., Wang, C., and Jiang, G. "Leveraging genetic reports and electronic health records for the prediction of primary cancers: Algorithm development and validation study". In: *JMIR Medical Informatics* 9.5 (2021), pp. 1–18. ISSN: 22919694. DOI: 10.2196/23586.

[69] Hutchings, R. "Questions of trust? Exploring the national data opt-out rate". In: *Nuffield Trust* (2023). URL: https://www.nuffieldtrust.org.uk/news-item/questions-of-trust-exploring-the-national-data-opt-out-rate.

[70] Tazare, J., Henderson, A. D., Morley, J., Blake, H. A., McDonald, H. I., Williamson, E. J., and Strongman, H. "NHS national data opt-outs: trends and potential consequences for health data research". In: *BJGP open* 8.3 (2024).

[71] Lee, D., Jiang, X., and Yu, H. "Harmonized representation learning on dynamic EHR graphs". In: *Journal of Biomedical Informatics* 106.November 2019 (2020), p. 103426. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103426. URL: https://doi.org/10.1016/j.jbi.2020.103426.

[72] Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., and Bauer, C. R. "Characterizing and managing missing structured data in electronic health records: Data analysis". In: *JMIR Medical Informatics* 6.1 (2018), pp. 1–12. ISSN: 22919694. DOI: 10.2196/medinform.8960.

[73] Groenwold, R. H. H. "Informative missingness in electronic health record systems: the curse of knowing". In: *Diagnostic and Prognostic Research* 4.1 (2020), pp. 4–9. DOI: 10.1186/s41512-020-00077-0.

[74] Folino, F., Pizzuti, C., and Ventura, M. "A comorbidity network approach to predict disease risk". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6266 LNCS.i (2010), pp. 102–109. ISSN: 03029743. DOI: 10.1007/978-3-642-15020-3{_}10.

[75] Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., and Connolly, T. M. "A systematic review of technologies and standards used in the development of rule-based clinical decision support systems". In: *Health and Technology* 12.4 (July 2022), pp. 713–727. ISSN: 2190-7188. DOI: 10.1007/s12553-022-00672-9. URL: https://link.springer.com/10.1007/s12553-022-00672-9.

[76] Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., and Ramkumar, P. N. "Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions". In: *Current Reviews in Musculoskeletal Medicine* 13.1 (Feb. 2020), pp. 69–76. ISSN: 1935-9748. DOI: 10.1007/s12178-020-09600-8. URL: http://link.springer.com/10.1007/s12178-020-09600-8.

[77] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models". In: *Journal of Clinical Epidemiology*

110 (2019), pp. 12–22. ISSN: 18785921. DOI: 10.1016/j.jclinepi.2019.02.004. URL: https://doi.org/10.1016/j.jclinepi.2019.02.004.

[78] Song, X., Liu, X., Liu, F., and Wang, C. "Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis". In: *International Journal of Medical Informatics* 151 (July 2021), p. 104484. ISSN: 13865056. DOI: 10.1016/j.ijmedinf.2021.104484. URL: https://linkinghub.elsevier.com/retrieve/pii/S1386505621001106.

[79] Ottenbacher, K. J., Linn, R. T., Smith, P. M., Illig, S. B., Mancuso, M., and Granger, C. V. "Comparison of logistic regression and neural network analysis applied to predicting living setting after hip fracture". In: *Annals of Epidemiology* 14.8 (2004), pp. 551–559. ISSN: 10472797. DOI: 10.1016/j.annepidem.2003.10.005.

[80] Joseph, G. B., McCulloch, C. E., Sohn, J. H., Pedoia, V., Majumdar, S., and Link, T. M. "AI MSK clinical applications: cartilage and osteoarthritis". In: *Skeletal Radiology* 51.2 (2022), pp. 331–343. ISSN: 14322161. DOI: 10.1007/s00256-021-03909-2. URL: https://doi.org/10.1007/s00256-021-03909-2.

[81] Worster, A., Fan, J., and Ismaila, A. "Understanding linear and logistic regression analyses". In: *CJEM* 9.02 (Mar. 2007), pp. 111–113. ISSN: 1481-8035. DOI: 10.1017/S1481803500014883. URL: https://www.cambridge.org/core/product/identifier/S1481803500014883/type/journal_article.

[82] Chan, D. X. H., Sim, Y. E., Chan, Y. H., Poopalalingam, R., and Abdullah, H. R. "Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: A single-center retrospective study". In: *BMJ Open* 8.3 (2018), pp. 1–11. ISSN: 20446055. DOI: 10.1136/bmjopen-2017-019427.

[83] Becker, T., Rousseau, A.-J., Geubbelmans, M., Burzykowski, T., and Valkenborg, D. "Decision trees and random forests". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 164.6 (Dec. 2023), pp. 894–897. ISSN: 08895406. DOI: 10.1016/j.ajodo.2023.09.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0889540623005188.

[84] Mustafa Abdullah, D. and Mohsin Abdulazeez, A. "Machine Learning Applications based on SVM Classification A Review". In: *Qubahan Academic Journal* 1.2 (Apr. 2021), pp. 81–90. ISSN: 2709-8206. DOI: 10.48161/qaj.v1n2a50. URL: https://journal.qubahan.com/index.php/qaj/article/view/50.

[85] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. "A review of clustering techniques and developments". In: *Neurocomputing* 267 (Dec. 2017), pp. 664–681. ISSN: 09252312. DOI: 10.1016/j.neucom.2017.06.053. URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231217311815.

[86] Larvin, H., Kang, J., Aggarwal, V. R., Pavitt, S., and Wu, J. *Systemic Multimorbidity Clusters in People with Periodontitis*. 2022. DOI: 10.1177/00220345221098910.

[87] Ningrum, D. N. A., Kung, W. M., Tzeng, I. S., Yuan, S. P., Wu, C. C., Huang, C. Y., Muhtar, M. S., Nguyen, P. A., Li, J. Y. C., and Wang, Y. C. "A deep learning model to predict knee osteoarthritis based on nonimage longitudinal medical record". In: *Journal of Multidisciplinary Healthcare* 14.June (2021), pp. 2477–2485. ISSN: 11782390. DOI: 10.2147/JMDH.S325179.

[88] Hanga, K. M., Kovalchuk, Y., and Gaber, M. M. "A graph-based approach to interpreting recurrent neural networks in process mining". In: *IEEE Access* 8 (2020), pp. 172923–172938. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3025999.

[89] Hochreiter, S. and Schmidhuber, J. "Long Short-Term Memory". In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://direct.mit.edu/neco/article/9/8/1735-1780/6109.

[90] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. "A ConvNet for the 2020s". In: (Jan. 2022). URL: http://arxiv.org/abs/2201.03545.

[91] Heckerman David Wellman, M. P. "Bayesian networks". In: *Communications of the ACM* 38.3 (1995), pp. 27–31.

[92] Arulkumaran, K., Deisenroth, M. P., Brundage, M., and Bharath, A. A. "Deep Reinforcement Learning: A Brief Survey". In: *IEEE Signal Processing Magazine* 34.6 (Nov.

2017), pp. 26–38. ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2743240. URL: http://ieeexplore.ieee.org/document/8103164/.

[93] Halilaj, E., Le, Y., Hicks, J., Hastie, T., and Delp, S. "Modeling and predicting osteoarthritis progression: data from the osteoarthritis initiative". In: *Osteoarthritis and Cartilage* 26.12 (Dec. 2018), pp. 1643–1650. ISSN: 10634584. DOI: 10.1016/j.joca.2018.08.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1063458418314158.

[94] Tiulpin, A., Klein, S., Bierma-Zeinstra, S. M. A., Thevenot, J., Rahtu, E., Meurs, J. v., Oei, E. H. G., and Saarakkala, S. "Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data". In: *Scientific Reports* 9.1 (Dec. 2019), p. 20038. ISSN: 2045-2322. DOI: 10.1038/s41598-019-56527-3. URL: http://www.nature.com/articles/s41598-019-56527-3.

[95] Burgess, R., Mansell, G., Bishop, A., Lewis, M., and Hill, J. "Predictors of functional outcome in musculoskeletal healthcare: An umbrella review". In: *European Journal of Pain* 24.1 (Jan. 2020), pp. 51–70. ISSN: 1090-3801. DOI: 10.1002/ejp.1483. URL: https://onlinelibrary.wiley.com/doi/10.1002/ejp.1483.

[96] Walsh, M. E., French, H. P., Wallace, E., Madden, S., King, P., Fahey, T., and Galvin, R. "Existing validated clinical prediction rules for predicting response to physiotherapy interventions for musculoskeletal conditions have limited clinical value: A systematic review". In: *Journal of Clinical Epidemiology* 135 (July 2021), pp. 90–102. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2021.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435621000421.

[97] Munoz-gama, J., Martin, N., and Fernandez-llatas, C. "Process Mining for Healthcare : Characteristics and Challenges". In: (2022).

[98] Theis, J., Galanter, W. L., Boyd, A. D., and Darabi, H. "Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture". In: *IEEE Journal of Biomedical and Health Informatics* 26.1 (2022), pp. 388–399. ISSN: 21682208. DOI: 10.1109/JBHI.2021.3092969.

[99] Farid, N. F., De Kamps, M., and Johnson, O. A. "Process mining in frail elderly care: A literature review". In: *HEALTHINF 2019 - 12th International Conference on Health Informatics, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019* (2019), pp. 332–339. DOI: 10.5220/0007392903320339.

[100] Conca, T., Saint-Pierre, C., Herskovic, V., Sepúlveda, M., Capurro, D., Prieto, F., and Fernandez-Llatas, C. "Multidisciplinary collaboration in the treatment of patients with type 2 diabetes in primary care: Analysis using process mining". In: *Journal of Medical Internet Research* 20.4 (2018). ISSN: 14388871. DOI: 10.2196/jmir.8884.

[101] Kurniati, A. P., Johnson, O., Hogg, D., and Hall, G. "Process Mining in Oncology : a Literature Review . White Rose Research Online URL for this paper : Version : Accepted Version (2016) Process Mining in Oncology : a Literature Review . In : Proceedings of the 6th Internationa". In: (2016).

[102] Khedkar, S., Gandhi, P., Shinde, G., and Subramanian, V. "Deep Learning and Explainable AI in Healthcare Using EHR". In: 2020, pp. 129–148. DOI: 10.1007/978-3-030-33966-1{_}7. URL: http://link.springer.com/10.1007/978-3-030-33966-1_7.

[103] Xin Teoh, Y., Othmani, A., Li Goh, S., Usman, J., and Lai, K. W. "Deciphering Knee Osteoarthritis Diagnostic Features With Explainable Artificial Intelligence: A Systematic Review". In: *IEEE Access* 12 (2024), pp. 109080–109108. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3439096. URL: https://ieeexplore.ieee.org/document/10623535/.

[104] M. Thimoteo, L., M. Vellasco, M., M. do Amaral, J., Figueiredo, K., Lie Yokoyama, C., and Marques, E. "Interpretable Machine Learning for COVID-19 Diagnosis Through Clinical Variables". In: *Anais do Congresso Brasileiro de Automática 2020*. sbabra, Dec. 2020. DOI: 10.48011/asba.v2i1.1590. URL: https://www.sba.org.br/open_journal_systems/index.php/sba/article/view/1590.

[105] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Damen, J. A., Debray, T. P., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Andaur Navarro, C. L., Reitsma, J. B., Sergeant, J. C., Shi, C., Skoetz, N., Smits, L. J., Snell,

K. I., Sperrin, M., Spijker, R., Steyerberg, E. W., Takada, T., Van Kuijk, S. M., Van Royen, F. S., Wallisch, C., Hooft, L., Moons, K. G., and Van Smeden, M. "Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal". In: *The BMJ* 369 (2020). ISSN: 17561833. DOI: 10.1136/bmj.m1328.

[106] Hossain, M. E., Khan, A., Moni, M. A., and Uddin, S. "Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.2 (2021), pp. 745–758. ISSN: 15579964. DOI: 10.1109/TCBB.2019.2937862.

[107] Carrasco-Ribelles, L. A., Llanes-Jurado, J., Gallego-Moll, C., Cabrera-Bean, M., Monteagudo-Zaragoza, M., Violán, C., and Zabaleta-del-Olmo, E. "Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review". In: *Journal of the American Medical Informatics Association* 30.12 (Nov. 2023), pp. 2072–2082. ISSN: 1067-5027. DOI: 10.1093/jamia/ocad168. URL: https://academic.oup.com/jamia/article/30/12/2072/7259105.

[108] Amirahmadi, A., Ohlsson, M., and Etminani, K. "Deep learning prediction models based on EHR trajectories: A systematic review". In: *Journal of Biomedical Informatics* 144 (Aug. 2023), p. 104430. ISSN: 15320464. DOI: 10.1016/j.jbi.2023.104430. URL: https://linkinghub.elsevier.com/retrieve/pii/S153204642300151X.

[109] Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review". In: *Journal of the American Medical Informatics Association* 24.1 (2017), pp. 198–208. ISSN: 1527974X. DOI: 10.1093/jamia/ocw042.

[110] Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J. "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in Neural Information Processing Systems* Nips (2016), pp. 3512–3520. ISSN: 10495258.

[111] Wu, T., Wang, Y., Wang, Y., Zhao, E., and Wang, G. "OA-MedSQL: Order-Aware Medical Sequence Learning for Clinical Outcome Prediction". In: *2021 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2022, pp. 1585–1589. ISBN: 9781665401265. DOI: 10.1109/bibm52615.2021.9669367.

[112] Clegg, A., Bates, C., Young, J., Ryan, R., Nichols, L., Ann Teale, E., Mohammed, M. A., Parry, J., and Marshall, T. "Development and validation of an electronic frailty index using routine primary care electronic health record data". In: *Age and Ageing* 45.3 (May 2016), pp. 353–360. ISSN: 0002-0729. DOI: 10.1093/ageing/afw039. URL: https://academic.oup.com/ageing/article-lookup/doi/10.1093/ageing/afw039.

[113] Khanna, A. K., Motamedi, V., Bouldin, B., Harwood, T., Pajewski, N. M., Saha, A. K., and Segal, S. "Automated Electronic Frailty Index–Identified Frailty Status and Associated Postsurgical Adverse Events". In: *JAMA Network Open* 6.11 (Nov. 2023), e2341915. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2023.41915. URL: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2811392.

[114] Boyd, P. J., Nevard, M., Ford, J. A., Khondoker, M., Cross, J. L., and Fox, C. "The electronic frailty index as an indicator of community healthcare service utilisation in the older population". In: *Age and Ageing* 48.2 (Mar. 2019), pp. 273–277. ISSN: 0002-0729. DOI: 10.1093/ageing/afy181. URL: https://academic.oup.com/ageing/article/48/2/273/5262695.

[115] Archer, L., Koshiaris, C., Lay-Flurrie, S., Snell, K. I. E., Riley, R. D., Stevens, R., Banerjee, A., Usher-Smith, J. A., Clegg, A., Payne, R. A., Hobbs, F. D. R., McManus, R. J., and Sheppard, J. P. "Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: retrospective cohort study". In: *BMJ* (Nov. 2022), e070918. ISSN: 1756-1833. DOI: 10.1136/bmj-2022-070918. URL: https://www.bmj.com/lookup/doi/10.1136/bmj-2022-070918.

[116] Gotz, D., Wang, F., and Perer, A. "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data". In: *Journal of Biomedical Informatics* 48 (2014), pp. 148–159. ISSN: 15320464. DOI: 10.1016/j.jbi.2014.01.007. URL: http://dx.doi.org/10.1016/j.jbi.2014.01.007.

[117] Perer, A. and Wang, F. "Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences". In: (2014), pp. 153–162. DOI: 10.1145/2557500.2557508.

[118] Liu, C., Wang, F., Hu, J., and Xiong, H. "Temporal phenotyping from longitudinal Electronic Health Records: A graph based framework". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015-Augus (2015), pp. 705–714. DOI: 10.1145/2783258.2783352.

[119] Ainsworth, J. and Buchan, I. "Combining Health Data Uses to Ignite Health System Learning." In: *Methods of information in medicine* 54.6 (2015), pp. 479–87. ISSN: 2511-705X. DOI: 10.3414/ME15-01-0064. URL: http://www.ncbi.nlm.nih.gov/pubmed/26395036.

[120] Berge, C. *Graphs and Hypergraphs*. North-Holland Publishing Company, 1976. ISBN: 9780720404791. URL: http://compalg.inf.elte.hu/~tony/Oktatas/Algoritmusok-hatekonysaga/Berge-hypergraphs.pdf.

[121] Lo, S. and Lin, C. "WMR–A Graph-Based Algorithm for Friend Recommendation". In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, Dec. 2006, pp. 121–128. ISBN: 0-7695-2747-7. DOI: 10.1109/WI.2006.202. URL: http://ieeexplore.ieee.org/document/4061351/.

[122] Veselkov, K., Gonzalez, G., Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., Bronstein, M., and Laponogov, I. "Hyperfoods: Machine intelligent mapping of cancer-beating molecules in foods". In: *Scientific Reports* 9 (2019), pp. 1–12. ISSN: 20452322. DOI: 10.1038/s41598-019-45349-y. URL: http://dx.doi.org/10.1038/s41598-019-45349-y.

[123] Ju, W., Luo, X., Ma, Z., Yang, J., Deng, M., and Zhang, M. "GHNN: Graph Harmonic Neural Networks for semi-supervised graph-level classification". In: *Neural Networks* (2022). ISSN: 08936080. DOI: 10.1016/j.neunet.2022.03.018. URL: https://doi.org/10.1016/j.neunet.2022.03.018.

[124] Saiod, A. K., Greunen, D. van, and Veldsman, A. "Electronic Health Records: Benefits and Challenges for Data Quality". In: 2017, pp. 123–156. DOI: 10.1007/978-3-319-58280-1{_}6. URL: http://link.springer.com/10.1007/978-3-319-58280-1_6.

[125] Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. "GRAM: Graph-based attention model for healthcare representation learning". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Part F1296 (2017), pp. 787–795. DOI: 10.1145/3097983.3098126.

[126] Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. "KAME: Knowledge-based Attention Model for Diagnosis Prediction in Healthcare". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, Oct. 2018, pp. 743–752. ISBN: 9781450360142. DOI: 10.1145/3269206.3271701. URL: https://dl.acm.org/doi/10.1145/3269206.3271701.

[127] Holme, P. and Saramäki, J. "Temporal networks". In: *Physics Reports* 519.3 (Oct. 2012), pp. 97–125. ISSN: 03701573. DOI: 10.1016/j.physrep.2012.03.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0370157312000841.

[128] Michail, O. "An Introduction to Temporal Graphs: An Algorithmic Perspective". In: *Internet Mathematics* 12.4 (July 2016), pp. 239–280. ISSN: 1542-7951. DOI: 10.1080/15427951.2016.1177801. URL: http://www.internetmathematicsjournal.com/article/1606.

[129] Longa, A., Lachi, V., Santin, G., Bianchini, M., Lepri, B., Lio, P., Scarselli, F., and Passerini, A. "Graph Neural Networks for temporal graphs: State of the art, open challenges, and opportunities". In: (Feb. 2023). URL: http://arxiv.org/abs/2302.01018.

[130] Min, S., Gao, Z., Peng, J., Wang, L., Qin, K., and Fang, B. "STGSN — A Spatial–Temporal Graph Neural Network framework for time-evolving social networks". In: *Knowledge-Based Systems* 214 (Feb. 2021), p. 106746. ISSN: 09507051. DOI: 10.1016/j.knosys.2021.106746. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950705121000095.

[131] Yan, S., Xiong, Y., and Lin, D. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition". In: (Jan. 2018). URL: http://arxiv.org/abs/1801.07455.

[132] Wu, M., Li, C., Shen, Z., He, S., Tang, L., Zheng, J., Fang, Y., Li, K., Cheng, Y., Shi, Z., Sheng, G., Liu, Y., Zhu, J., Ye, X., Chen, J., Chen, W., Li, L., Sun, Y., and Chen, J. "Use of temporal contact graphs to understand the evolution of COVID-19 through contact tracing data". In: *Communications Physics* 5.1 (Nov. 2022), p. 270. ISSN: 2399-3650. DOI: 10.1038/s42005-022-01045-4. URL: https://www.nature.com/articles/s42005-022-01045-4.

[133] Longa, A., Lachi, V., Santin, G., Bianchini, M., Lepri, B., Lio, P., Scarselli, F., and Passerini, A. "Graph Neural Networks for temporal graphs: State of the art, open challenges, and opportunities". In: (Feb. 2023). URL: http://arxiv.org/abs/2302.01018.

[134] Schrodt, J., Dudchenko, A., Knaup-Gregori, P., and Ganzinger, M. "Graph-Representation of Patient Data: a Systematic Literature Review". In: *Journal of Medical Systems* 44.4 (2020), p. 7. ISSN: 1573689X. DOI: 10.1007/s10916-020-1538-4.

[135] Wardle, M. T., Reavis, K. M., and Snowden, J. M. "Measurement error and information bias in causal diagrams: mapping epidemiological concepts and graphical structures". In: *International journal of epidemiology* 53.6 (2024), dyae141.

[136] Liu, X., Wang, H., He, T., Liao, Y., and Jian, C. "Recent Advances in Representation Learning for Electronic Health Records: A Systematic Review". In: *Journal of Physics: Conference Series* 2188.1 (2022), p. 012007. ISSN: 1742-6588. DOI: 10.1088/1742-6596/2188/1/012007.

[137] Hamilton, W. L. "Graph Representation Learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14.3 (Sept. 2020), pp. 1–159. ISSN: 1939-4608. DOI: 10.2200/S01045ED1V01Y202009AIM046. URL: https://www.morganclaypool.com/doi/10.2200/S01045ED1V01Y202009AIM046.

[138] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. "The graph neural network model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. ISSN: 10459227. DOI: 10.1109/TNN.2008.2005605.

[139] Khan, A., Uddin, S., and Srinivasan, U. "Adapting graph theory and social network measures on healthcare data - A new framework to understand chronic disease pro-

gression". In: *ACM International Conference Proceeding Series* February (2016). DOI: 10.1145/2843043.2843380.

[140] Khan, A., Uddin, S., and Srinivasan, U. "Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes". In: *Expert Systems with Applications* 136 (Dec. 2019), pp. 230–241. ISSN: 09574174. DOI: 10.1016/j.eswa.2019.05.048.

[141] Wanyan, T., Honarvar, H., Azad, A., Ding, Y., and Glicksberg, B. S. "Deep learning with heterogeneous graph embeddings for mortality prediction from electronic health records". In: *Data Intelligence* 3.3 (2021), pp. 329–339. ISSN: 2641435X. DOI: 10.1162/dint{_}a{_}00097.

[142] Yao, H. R., Chang, D. C., Frieder, O., Huang, W., Liang, I. C., and Hung, C. F. "Cross-Global Attention Graph Kernel Network Prediction of Drug Prescription". In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2020*. Association for Computing Machinery, Inc, Sept. 2020. ISBN: 9781450379649. DOI: 10.1145/3388440.3412459.

[143] Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., and Liu, N. "Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies". In: *Journal of Biomedical Informatics* 126.July 2021 (2022), p. 103980. ISSN: 15320464. DOI: 10.1016/j.jbi.2021.103980. URL: https://doi.org/10.1016/j.jbi.2021.103980.

[144] Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., and Roberts, K. "Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review". In: *Journal of Biomedical Informatics* 115.October 2020 (2021), p. 103671. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103671. URL: https://doi.org/10.1016/j.jbi.2020.103671.

[145] Choi, E., Xu, Z., Li, Y., Dusenberry, M. W., Flores, G., Xue, E., and Dai, A. M. "Learning the graphical structure of electronic health records with graph convolutional transformer". In: *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* (2020), pp. 606–613. ISSN: 2159-5399. DOI: 10.1609/aaai.v34i01.5400.

[146] Yao, H. R., Chang, D. C., Frieder, O., Huang, W., and Lee, T. S. "Graph Kernel Prediction of Drug Prescription". In: *2019 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings* (2019). DOI: 10.1109/BHI.2019.8834676.

[147] Zhang, J., Gong, J., and Barnes, L. "HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records". In: *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017* (2017), pp. 214–221. DOI: 10.1109/CHASE.2017.80.

[148] Zhang, S., Liu, L., Li, H., Xiao, Z., and Cui, L. "MTPGraph: A data-driven approach to predict medical risk based on temporal profile graph". In: *Proceedings - 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 10th IEEE International Conference on Big Data Science and Engineering and 14th IEEE International Symposium on Parallel and Distributed Proce* 1 (2016), pp. 1174–1181. DOI: 10.1109/TrustCom.2016.0191.

[149] Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., Estarli, M., Barrera, E. S., Martínez-Rodríguez, R., Baladia, E., Agüero, S. D., Camacho, S., Buhring, K., Herrero-López, A., Gil-González, D. M., Altman, D. G., Booth, A., Chan, A. W., Chang, S., Clifford, T., Dickersin, K., Egger, M., Gøtzsche, P. C., Grimshaw, J. M., Groves, T., Helfand, M., Higgins, J., Lasserson, T., Lau, J., Lohr, K., McGowan, J., Mulrow, C., Norton, M., Page, M., Sampson, M., Schünemann, H., Simera, I., Summerskill, W., Tetzlaff, J., Trikalinos, T. A., Tovey, D., Turner, L., and Whitlock, E. "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement". In: *Revista Espanola de Nutricion Humana y Dietetica* 20.2 (2016), pp. 148–160. ISSN: 21731292. DOI: 10.1186/2046-4053-4-1.

[150] Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. "Rayyan-a web and mobile app for systematic reviews". In: *Systematic Reviews* 5.1 (2016), pp. 1–10. ISSN: 20464053. DOI: 10.1186/s13643-016-0384-4. URL: http://dx.doi.org/10.1186/s13643-016-0384-4.

[151] Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., and Mallett, S. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies". In: *Annals of Internal Medicine* 170.1 (Jan. 2019), p. 51. ISSN: 0003-4819. DOI: 10.7326/M18-1376. URL: http://annals.org/article.aspx?doi=10.7326/M18-1376.

[152] Moons, K. G., Groot, J. A. de, Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., and Collins, G. S. "Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist". In: *PLoS Medicine* 11.10 (2014). ISSN: 15491676. DOI: 10.1371/journal.pmed.1001744. URL: www.plosmedicine.org.

[153] McCormick, T. H., Cynthia, R., and Madigan, D. "a Hierarchical Model for Association Rule Mining of Sequential Events : an Approach To". In: (2000), pp. 1–19.

[154] Loscalzo, J., Kohane, I., and Barabasi, A. L. "Human disease classification in the postgenomic era: A complex systems approach to human pathobiology". In: *Molecular Systems Biology* 3.124 (2007). ISSN: 17444292. DOI: 10.1038/msb4100163.

[155] Baglioni, M., Pieroni, S., Geraci, F., Mariani, F., Molinaro, S., Pellegrini, M., and Lastres, E. "A new framework for distilling higher quality information from health data via social network analysis". In: *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013* (2013), pp. 48–55. DOI: 10.1109/ICDMW.2013.142.

[156] Chang, C. D., Wang, C. C., and Jiang, B. C. "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors". In: *Expert Systems with Applications* 38.5 (2011), pp. 5507–5513. ISSN: 09574174. DOI: 10.1016/j.eswa.2010.10.086. URL: http://dx.doi.org/10.1016/j.eswa.2010.10.086.

[157] Haddaway, N. R., Page, M. J., Pritchard, C. C., and McGuinness, L. A. "PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis". In: *Campbell Systematic Reviews* 18.2 (2022), e1230.

[158] Zhang, S., Liu, L., Li, H., and Cui, L. "Collaborative Prediction Model of Disease Risk by Mining Electronic Health Records". In: *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*. Vol. 201. 2017, pp. 71–82. ISBN: 9783319592886. DOI: 10.1007/978-3-319-59288-6{_}7. URL: http://link.springer.com/10.1007/978-3-319-59288-6_7.

[159] Golmaei, S. N. and Luo, X. "DeepNote-GNN: Predicting hospital readmission using clinical notes and patient network". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2021*. Association for Computing Machinery, Inc, Jan. 2021. ISBN: 9781450384506. DOI: 10.1145/3459930.3469547.

[160] Sun, C., Dui, H., and Li, H. "Interpretable time-aware and co-occurrence-aware network for medical prediction". In: *BMC Medical Informatics and Decision Making* 21.1 (2021), pp. 1–12. ISSN: 14726947. DOI: 10.1186/s12911-021-01662-z. URL: https://doi.org/10.1186/s12911-021-01662-z.

[161] Qian, Z., Alaa, A. M., Bellot, A., Rashbass, J., and Schaar, M. van der. "Learning Dynamic and Personalized Comorbidity Networks from Event Data using Deep Diffusion Processes". In: 108 (2020). URL: http://arxiv.org/abs/2001.02585.

[162] Hettige, B., Wang, W., Li, Y. F., Le, S., and Buntine, W. "MedGraph: Structural and temporal representation learning of electronic medical records". In: *Frontiers in Artificial Intelligence and Applications* 325 (2020), pp. 1810–1817. ISSN: 09226389. DOI: 10.3233/FAIA200296.

[163] Chen, L., Li, X., Sheng, Q. Z., Peng, W. C., Bennett, J., Hu, H. Y., and Huang, N. "Mining Health Examination Records - A Graph-Based Approach". In: *IEEE Transactions on Knowledge and Data Engineering* 28.9 (2016), pp. 2423–2437. ISSN: 10414347. DOI: 10.1109/TKDE.2016.2561278.

[164] Yao, H. R., Chang, D. C., Frieder, O., Huang, W., and Lee, T. S. "Multiple graph kernel fusion prediction of drug prescription". In: *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (2019), pp. 103–112. DOI: 10.1145/3307339.3342134.

[165] Xue, Y., Klabjan, D., and Luo, Y. "Predicting ICU readmission using grouped physiological and medication trends". In: *Artificial Intelligence in Medicine* 95.3 (Apr. 2019), pp. 27–37. ISSN: 09333657. DOI: 10.1016/j.artmed.2018.08.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365717306486.

[166] Lu, Q., Nguyen, T. H., and Dou, D. "Predicting Patient Readmission Risk from Medical Text via Knowledge Graph Enhanced Multiview Graph Convolution". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, July 2021, pp. 1990–1994. ISBN: 9781450380379. DOI: 10.1145/3404835.3463062. URL: https://dl.acm.org/doi/10.1145/3404835.3463062.

[167] Liu, X., Wang, H., He, T., and Gong, X. "Research on Intelligent Diagnosis Model of Electronic Medical Record Based on Graph Transformer". In: *Proceedings - 2021 6th International Conference on Computational Intelligence and Applications, ICCIA 2021* (2021), pp. 73–78. DOI: 10.1109/ICCIA52886.2021.00022.

[168] Kamkar, I., Gupta, S., Li, C., Phung, D., and Venkatesh, S. "Stable clinical prediction using graph support vector machines". In: *Proceedings - International Conference on Pattern Recognition* 0 (2016), pp. 3332–3337. ISSN: 10514651. DOI: 10.1109/ICPR.2016.7900148.

[169] Wang, S. and Liu, J. "TAGNet: Temporal Aware Graph Convolution Network for Clinical Information Extraction". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, Dec. 2020, pp. 2105–2108. ISBN: 978-1-7281-6215-7. DOI: 10.1109/BIBM49941.2020.9313530. URL: https://ieeexplore.ieee.org/document/9313530/.

[170] Chang, D. C., Frieder, O., Hung, C. F., and Yao, H. R. "The Analysis from Nonlinear Distance Metric to Kernel-based Drug Prescription Prediction System". In: (2021), pp. 1–21. URL: http://arxiv.org/abs/2102.02446.

[171] Kim, I. C. and Jung, Y. G. "Using Bayesian networks to analyze medical data". In: *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)* 2734 (2003), pp. 317–327. ISSN: 03029743. DOI: 10.1007/3-540-45065-3{_}28.

[172] Zhu, W. and Razavian, N. *Variationally regularized graph-based representation learning for electronic health records*. Vol. 1. 1. Association for Computing Machinery, 2021, pp. 1–13. ISBN: 9781450383592. DOI: 10.1145/3450439.3451855.

[173] Cho, H. N., Ahn, I., Gwon, H., Kang, H. J., Kim, Y., Seo, H., Choi, H., Kim, M., Han, J., Kee, G., Jun, T. J., and Kim, Y. H. "Heterogeneous graph construction and HinSAGE learning from electronic medical records". In: *Scientific Reports* 12.1 (2022), pp. 1–9. ISSN: 20452322. DOI: 10.1038/s41598-022-25693-2. URL: https://doi.org/10.1038/s41598-022-25693-2.

[174] Xu, Y., Ying, H., Qian, S., Zhuang, F., Zhang, X., Wang, D., Wu, J., and Xiong, H. "Time-Aware Context-Gated Graph Attention Network for Clinical Risk Prediction". In: *IEEE Transactions on Knowledge and Data Engineering* 35.7 (2023), pp. 7557–7568. ISSN: 15582191. DOI: 10.1109/TKDE.2022.3181780.

[175] OpenSAFELY. *About OpenSAFELY*. 2022. URL: https://www.opensafely.org/about/.

[176] Yang, C., Kors, J. A., Ioannou, S., John, L. H., Markus, A. F., Rekkas, A., De Ridder, M. A., Seinen, T. M., Williams, R. D., and Rijnbeek, P. R. "Trends in the conduct and reporting of clinical prediction model development and validation: A systematic review". In: *Journal of the American Medical Informatics Association* 29.5 (2022), pp. 983–989. ISSN: 1527974X. DOI: 10.1093/jamia/ocac002.

[177] Andaur Navarro, C. L., Damen, J. A., Smeden, M. van, Takada, T., Nijman, S. W., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G., and Hooft, L. "Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models". In: *Journal of Clinical Epidemiology* January (Nov. 2023), p. 113260. ISSN: 08954356. DOI: 10.1016/j.jclinepi.2022.11.015. URL: https://doi.org/10.1016/j.jclinepi.2022.11.015%20https://linkinghub.elsevier.com/retrieve/pii/S0895435622003006.

[178] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. "Data Descriptor: MIMIC-III, a freely accessible critical care database". In: *Sci. Data* 3:160035 (2016), pp. 1–9.

[179] Murray, R. E., Ryan, P. B., and Reisinger, S. J. "Design and Validation of a Data Simulation Model for Longitudinal Healthcare Data". In: *AMIA Annual Symposium Proceedings Archive* (2011), pp. 1176–1185.

[180] Gille, F. and Brall, C. "Limits of data anonymity: lack of public awareness risks trust in health system activities". In: *Life Sciences, Society and Policy* 17.1 (2021), pp. 1–8. ISSN: 21957819. DOI: 10.1186/s40504-021-00115-9.

[181] Goldacre, B. and Morley, J. *Better, Broader, Safer: Using Health Data for Research and Analysis*. Tech. rep. Department of Health and Social Care, 2022. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf?utm_campaign=846512_PRESS%20RELEASE%20Goldacre%20review&utm_medium=email&utm_source=NHS%20Confe.

[182] Gómez-Carmona, O., Casado-Mansilla, D., Kraemer, F. A., López-de-Ipiña, D., and García-Zubia, J. "Exploring the computational cost of machine learning at the edge for human-centric Internet of Things". In: *Future Generation Computer Systems* 112 (Nov. 2020), pp. 670–683. ISSN: 0167739X. DOI: 10.1016/j.future.2020.06.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X20304106.

[183] Steyerberg, E. W. and Vergouwe, Y. "Towards better clinical prediction models: seven steps for development and an ABCD for validation". In: *European heart journal* 35.29 (2014), pp. 1925–1931.

[184] Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., and Chao, L. S. "Learning deep transformer models for machine translation". In: *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 2020, pp. 1810–1822. ISBN: 9781950737482. DOI: 10.18653/v1/p19-1176. URL: https://github..

[185] Sun, Z., Harit, A., Yu, J., Cristea, A. I., and Shi, L. *A Brief Survey of Deep Learning Approaches for Learning Analytics on MOOCs*. Vol. 12677 LNCS. Springer International Publishing, 2021, pp. 28–37. ISBN: 9783030804206. DOI: 10.1007/978-3-030-80421-3{_}4. URL: http://dx.doi.org/10.1007/978-3-030-80421-3_4.

[186] Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement". In: *BMC Medicine* 13.1 (2015), pp. 1–10. ISSN: 17417015. DOI: 10.1186/s12916-014-0241-z.

[187] Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Van Calster, B., Smeden, M. van, Riley, R. D., and Moons, K. G. "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence". In: *BMJ Open* 11.7 (2021), pp. 1–7. ISSN: 20446055. DOI: 10.1136/bmjopen-2020-048008.

[188] Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G., Smeden, M. van, Sperrin, M., Bullock, G. S., and Riley, R. D. "Evaluation of clinical prediction models (part 1): from development to external validation". In: *Bmj* part 1 (2024). ISSN: 17561833. DOI: 10.1136/bmj-2023-074819.

[189] Riley, R. D., Archer, L., Snell, K. I., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., and Collins, G. S. "Evaluation of clinical prediction models (part 2): how to undertake an external validation study". In: *Bmj* part 2 (2024), pp. 1–12. ISSN: 17561833. DOI: 10.1136/bmj-2023-074820.

[190] Riley, R. D., Snell, K. I., Archer, L., Ensor, J., Debray, T. P., Van Calster, B., Smeden, M. van, and Collins, G. S. "Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study". In: *Bmj* part 3 (2024). ISSN: 17561833. DOI: 10.1136/bmj-2023-074821.

[191] Liu, K., Tatinati, S., and Khong, A. W. "A weighted feature extraction technique based on temporal accumulation of learner behavior features for early prediction of dropouts". In: *Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2020* (2020), pp. 295–302. DOI: 10.1109/TALE48869.2020.9368317.

[192] Wang, L. and Wang, H. "Learning behavior analysis and dropout rate prediction based on MOOCs data". In: *Proceedings - 10th International Conference on Information Technology in Medicine and Education, ITME 2019* (2019), pp. 419–423. DOI: 10.1109/ITME.2019.00100.

[193] Kumar, S., Zhang, X., and Leskovec, J. "Predicting Dynamic Embedding Trajectory in Temporal Interaction Networks". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY, USA: ACM, July 2019, pp. 1269–1278. ISBN: 9781450362016. DOI: 10.1145/3292500.3330895. URL: https://dl.acm.org/doi/10.1145/3292500.3330895.

[194] Haiyang, L., Wang, Z., Benachour, P., and Tubman, P. "A time series classification method for behaviour-based dropout prediction". In: *Proceedings - IEEE 18th International Conference on Advanced Learning Technologies, ICALT 2018* (2018), pp. 191–195. DOI: 10.1109/ICALT.2018.00052.

[195] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. "Temporal Graph Networks for Deep Learning on Dynamic Graphs". In: *arXiv preprint* (June 2020), pp. 1–16. URL: http://arxiv.org/abs/2006.10637.

[196] Chen, J., Liao, S., Hou, J., Wang, K., and Wen, J. "GST-GCN: A Geographic-Semantic-Temporal Graph Convolutional Network for Context-aware Traffic Flow Prediction on Graph Sequences". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Vol. 2020-Octob. IEEE, Oct. 2020, pp. 1604–1609. ISBN: 978-1-7281-8526-2. DOI: 10.1109/SMC42975.2020.9282828. URL: https://ieeexplore.ieee.org/document/9282828/.

[197] Edmond Meku Fotso, J., Batchakui, B., Nkambou, R., and Okereke, G. "Algorithms for the Development of Deep Learning Models for Classification and Prediction of Behaviour in MOOCS." In: *Proceedings of 2020 IEEE Learning With MOOCS, LWMOOCS 2020* (2020), pp. 180–184. DOI: 10.1109/LWMOOCS50143.2020.9234363.

[198] Geloven, N. v., Giardiello, D., Bonneville, E., Teece, L., Ramspek, C., Smeden, M. v., Snell, K., Calster, B. v., Pohar-Perme, M., Riley, R., Putter, H., and Steyerberg, E.

"Validation of prediction models in the presence of competing risks : a guide through modern methods". In: *British Medical Journal* (2022).

[199] Graham, B. "Sparse 3D convolutional neural networks". In: (2015), pp. 1–150. DOI: 10.5244/c.29.150.

[200] Wang, Y., Cai, Y., Liang, Y., and Ding, H. "Adaptive Data Augmentation on Temporal Graphs". In: *NeurIPS* NeurIPS (2021), pp. 1–13.

[201] Wang, Y., Cai, Y., Liang, Y., Ding, H., Wang, C., and Hooi, B. "Time-Aware Neighbor Sampling for Temporal Graph Networks". In: (2021), pp. 1–12. URL: http://arxiv.org/abs/2112.09845.

[202] Zhang, Y., Xiong, Y., Li, D., Shan, C., Ren, K., and Zhu, Y. "CoPE: Modeling Continuous Propagation and Evolution on Interaction Graph". In: *International Conference on Information and Knowledge Management, Proceedings* (2021), pp. 2627–2636. DOI: 10.1145/3459637.3482419.

[203] Zhou, Y., Luo, S., Pan, L., Liu, L., and Song, D. "Continuous temporal network embedding by modeling neighborhood propagation process". In: *Knowledge-Based Systems* 239 (2022), p. 107998. ISSN: 09507051. DOI: 10.1016/j.knosys.2021.107998. URL: https://doi.org/10.1016/j.knosys.2021.107998.

[204] Wen, Y., Tian, Y., Wen, B., Zhou, Q., Cai, G., and Liu, S. "Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs". In: *Tsinghua Science and Technology* 25.3 (2020), pp. 336–347. ISSN: 18787606. DOI: 10.26599/TST.2019.9010013.

[205] Ren, Y., Huang, S., and Zhou, Y. "Deep learning and integrated learning for predicting student's withdrawal behavior in MOOC". In: *Proceedings - 2021 2nd International Conference on Education, Knowledge and Information Management, ICEKIM 2021* (2021), pp. 81–84. DOI: 10.1109/ICEKIM52309.2021.00026.

[206] Zheng, Y., Gao, Z., Wang, Y., and Fu, Q. "MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series". In: *IEEE Access* 8 (2020), pp. 225324–225335. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3045157.

[207] Zhang, Y., Chang, L., and Liu, T. "MOOCs Dropout Prediction Based on Hybrid Deep Neural Network". In: *Proceedings - 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, CyberC 2020* (2020), pp. 197–203. DOI: 10.1109/CyberC49757.2020.00039.

[208] Hong, B., Wei, Z., and Yang, Y. "Discovering learning behavior patterns to predict dropout in MOOC". In: *2017 12th International Conference on Computer Science and Education (ICCSE)*. Iccse. IEEE, Aug. 2017, pp. 700–704. ISBN: 978-1-5090-2508-4. DOI: 10.1109/ICCSE.2017.8085583. URL: http://ieeexplore.ieee.org/document/8085583/.

[209] Cheng, D., Wang, X., Zhang, Y., and Zhang, L. "Graph Neural Network for Fraud Detection via Spatial-temporal Attention". In: *IEEE Transactions on Knowledge and Data Engineering* 14.8 (2020), pp. 1–1. ISSN: 1041-4347. DOI: 10.1109/TKDE.2020.3025588. URL: https://ieeexplore.ieee.org/document/9204584/.

[210] Doosti, B., Naha, S., Mirbagheri, M., and Crandall, D. "HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Mar. 2020), pp. 6607–6616. URL: https://arxiv.org/abs/2004.00060v1.

[211] Wu, Z., Wang, M., Wang, J., Zhang, W., Fang, M., and Xu, T. *DeepWORD: A GCN-based Approach for Owner-Member Relationship Detection in Autonomous Driving*. Institute of Electrical and Electronics Engineers (IEEE), Mar. 2021, pp. 1–6. URL: https://arxiv.org/abs/2103.16099v2.

[212] Pham Van, L. L., Bach Tran, Q., Pham, T. L., and Long Tran, Q. "Node-aware convolution in Graph Neural Networks for Predicting molecular properties". In: *Proceedings - 2020 12th International Conference on Knowledge and Systems Engineering, KSE 2020* (Nov. 2020), pp. 120–125. DOI: 10.1109/KSE50997.2020.9287744.

[213] Heuts, S., Kawczynski, M. J., Velders, B. J., Brophy, J. M., Hickey, G. L., and Kowalewski, M. "Statistical primer: an introduction into the principles of Bayesian statistical analyses in clinical trials". In: *European Journal of Cardio-Thoracic Surgery* 67.4 (2025), ezaf139.

[214] Chen, X., Tang, H., Lin, J., and Zeng, R. "Temporal trends in the disease burden of osteoarthritis from 1990 to 2019, and projections until 2030". In: *PLOS ONE* 18.7 (July 2023). Ed. by Sung, W.-W., e0288561. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0288561. URL: https://dx.plos.org/10.1371/journal.pone.0288561.

[215] Arslan, I. G., Damen, J., Wilde, M., Driest, J. J., Bindels, P. J., Lei, J., Schiphof, D., and Bierma-Zeinstra, S. M. "Incidence and prevalence of knee osteoarthritis using codified and narrative data from electronic health records: a population-based study". In: *Arthritis Care & Research* (2022), pp. 0–3. ISSN: 2151-464X. DOI: 10.1002/acr.24861.

[216] Keavy, R. "The prevalence of musculoskeletal presentations in general practice: an epidemiological study". In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 70 (2020), pp. 1–7. ISSN: 14785242. DOI: 10.3399/bjgp20X711497.

[217] Government, M. o. H. C. and Local. *The English Indices of Deprivation 2019 - Technical Report*. Tech. rep. 2019. URL: https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-%20technical-report.

[218] Dhiman, P., Ma, J., Andaur Navarro, C. L., Speich, B., Bullock, G., Damen, J. A., Hooft, L., Kirtley, S., Riley, R. D., Van Calster, B., Moons, K. G., and Collins, G. S. "Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review". In: *BMC Medical Research Methodology* 22.1 (2022), pp. 1–16. ISSN: 14712288. DOI: 10.1186/s12874-022-01577-x. URL: https://doi.org/10.1186/s12874-022-01577-x.

[219] Kingsbury, S. R., Smith, L. K., Czoski Murray, C. J., Pinedo-Villanueva, R., Judge, A., West, R., Smith, C., Wright, J. M., Arden, N. K., Thomas, C. M., Kolovos, S., Shuweihdi, F., Garriga, C., Bitanihirwe, B. K., Hill, K., Matu, J., Stone, M., and Conaghan, P. G. "Safety of disinvestment in mid-to late-term follow-up post primary hip and knee replacement: the UK SAFE evidence synthesis and recommendations". In: *Health and Social Care Delivery Research* 10.16 (2022). ISSN: 27550079. DOI: 10.3310/KODQ0769.

[220] Leung, K., Zhang, B., Tan, J., Shen, Y., Geras, K. J., Babb, J. S., Cho, K., Chang, G., and Deniz, C. M. "Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis

by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative". In: *Radiology* 296.3 (Sept. 2020), pp. 584–593. ISSN: 0033-8419. DOI: 10.1148/radiol.2020192091. URL: http://pubs.rsna.org/doi/10.1148/radiol.2020192091.

[221] Xu, Y., Hao, X., Weixuan, L., Liuu, H., Guo, J., Wang, W., Ruan, H., Sun, Z., and Fan, C. "Development and Validation of a Deep-Learning Model to Predict Total Hip Replacement on Radiographs". In: *The Journal of Bone and Joint Surgery* 106.5 (2024), pp. 389–396. URL: https://journals.lww.com/jbjsjournal/abstract/2024/03060/development_and_validation_of_a_deep_learning.2.aspx.

[222] Sun, C., Hong, S., Song, M., and Li, H. "A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data". In: (2020), pp. 1–20. URL: http://arxiv.org/abs/2010.12493.

[223] Pulikottil, S. C. and Gupta, M. "ONet - A Temporal Meta Embedding Network for MOOC Dropout Prediction". In: *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020* (2020), pp. 5209–5217. DOI: 10.1109/BigData50022.2020.9378001.

[224] Barry, N., Kendrick, J., Molin, K., Li, S., Rowshanfarzad, P., Hassan, G. M., Dowling, J., Parizel, P. M., Hofman, M. S., and Ebert, M. A. "Evaluating the impact of the Radiomics Quality Score: a systematic review and meta-analysis". In: *European Radiology* (2025), pp. 1–13.

[225] Jenkins, P. J., Clement, N. D., Hamilton, D. F., Gaston, P., Patton, J. T., and Howie, C. R. "Predicting the cost-effectiveness of total hip and knee replacement". In: *The Bone & Joint Journal* 95-B.1 (Jan. 2013), pp. 115–121. ISSN: 2049-4394. DOI: 10.1302/0301-620X.95B1.29835. URL: https://online.boneandjoint.org.uk/doi/10.1302/0301-620X.95B1.29835.

[226] Arthritis and (ARMA), M. A. *Musculoskeletal Health Inequalities and Deprivation*. ARMA, 2024. URL: https://arma.uk.net/wp-content/uploads/2024/08/Musculoskeletal-Health-Inequalities-and-Deprivation-report_v08-SMALL.pdf.

[227] Goodman, S. M., Mannstadt, I., Gibbons, J. A. B., Rajan, M., Bass, A., Russell, L., Mehta, B., Figgie, M., Parks, M. L., Venkatachalam, S., et al. "Healthcare disparities: patients' perspectives on barriers to joint replacement". In: *BMC Musculoskeletal Disorders* 24.1 (2023), p. 976.

[228] Sperrin, M., Riley, R. D., Collins, G. S., and Martin, G. P. "Targeted validation: validating clinical prediction models in their intended population and setting". In: *Diagnostic and Prognostic Research* 6.1 (Dec. 2022), p. 24. ISSN: 2397-7523. DOI: 10.1186/s41512-022-00136-8. URL: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-022-00136-8.

[229] Yuniar Purbasari, I., Priharyoto Bayuseno, A., Isnanto, R. R., Indah Winarni, T., and Jamari, J. "Artificial Intelligence and Machine Learning in Prediction of Total Hip Arthroplasty Outcome: A Bibliographic Review". In: *E3S Web of Conferences* 448 (Nov. 2023). Ed. by Isnanto, R., Hadiyanto, and Warsito, B., p. 02054. ISSN: 2267-1242. DOI: 10.1051/e3sconf/202344802054. URL: https://www.e3s-conferences.org/10.1051/e3sconf/202344802054.

[230] Riley, R. D., Ensor, J., Snell, K. I. E., Debray, T. P. A., Altman, D. G., Moons, K. G. M., and Collins, G. S. "External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges". In: *BMJ* (June 2016), p. i3140. ISSN: 1756-1833. DOI: 10.1136/bmj.i3140. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.i3140.

[231] Neufeld, M. E. and Masri, B. A. "Can the Oxford Knee and Hip Score identify patients who do not require total knee or hip arthroplasty?" In: *The Bone & Joint Journal* 101-B.6_Supple_B (June 2019), pp. 23–30. ISSN: 2049-4394. DOI: 10.1302/0301-620X.101B6.BJJ-2018-1460.R1. URL: https://online.boneandjoint.org.uk/doi/10.1302/0301-620X.101B6.BJJ-2018-1460.R1.

[232] Thuraisingam, S., Chondros, P., Manski-Nankervis, J.-A., Spelman, T., Choong, P. F., Gunn, J., and Dowsey, M. M. "Developing and internally validating a prediction model for total knee replacement surgery in patients with osteoarthritis". In: *Osteoarthritis and Cartilage Open* 4.3 (Sept. 2022), p. 100281. ISSN: 26659131. DOI: 10.1016/j.

ocarto.2022.100281. URL: https://linkinghub.elsevier.com/retrieve/pii/S2665913122000498.

[233] Yu, B., Li, M., Zhang, J., and Zhu, Z. "3D graph convolutional networks with temporal graphs: A spatial information free framework for traffic forecasting". In: *arXiv* (2019). ISSN: 23318422.

[234] Liu, N., Feng, Q., and Hu, X. "Interpretability in Graph Neural Networks". In: *Graph Neural Networks: Foundations, Frontiers, and Applications*. Singapore: Springer Nature Singapore, 2022, pp. 121–147. DOI: 10.1007/978-981-16-6054-2{_}7. URL: https://link.springer.com/10.1007/978-981-16-6054-2_7.

[235] Sheedy, A. N., Wactawski-Wende, J., Hovey, K. M., and LaMonte, M. J. "Discontinuation of hormone therapy and bone mineral density: does physical activity modify that relationship?" In: *Menopause* 30.12 (Dec. 2023), pp. 1199–1205. ISSN: 1530-0374. DOI: 10.1097/GME.0000000000002272. URL: https://journals.lww.com/10.1097/GME.0000000000002272.

[236] Rizzoli, R. and Bonjour, J.-P. "Hormones and bones". In: *The Lancet* 349 (Mar. 1997), S20–S23. ISSN: 01406736. DOI: 10.1016/S0140-6736(97)90007-6. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673697900076.

[237] Gopfert, A., Deeny, S. R., Fisher, R., and Stafford, M. "Primary care consultation length by deprivation and multimorbidity in England: an observational study using electronic patient records". In: *British Journal of General Practice* 71.704 (Mar. 2021), e185–e192. ISSN: 0960-1643. DOI: 10.3399/bjgp20X714029. URL: http://bjgp.org/lookup/doi/10.3399/bjgp20X714029.

[238] Getzen, E., Ungar, L., Mowery, D., Jiang, X., and Long, Q. "Mining for equitable health: Assessing the impact of missing data in electronic health records". In: *Journal of Biomedical Informatics* 139 (Mar. 2023), p. 104269. ISSN: 15320464. DOI: 10.1016/j.jbi.2022.104269. URL: https://linkinghub.elsevier.com/retrieve/pii/S153204642200274X.

[239] Carriero, A., Luijken, K., Hond, A. de, Moons, K. G., Calster, B. van, and Smeden, M. van. "The harms of class imbalance corrections for machine learning based prediction models: a simulation study". In: (2024). URL: http://arxiv.org/abs/2404.19494.

[240] Vinciguerra, C., Gueguen, A., Revel, M., Heuleu, J. N., Amor, B., and Dougados, M. "Predictors of the need for total hip replacement in patients with osteoarthritis of the hip." In: *Revue du rhumatisme (English ed.)* 62.9 (Oct. 1995), pp. 563–70. ISSN: 1169-8446. URL: http://www.ncbi.nlm.nih.gov/pubmed/8574628.

[241] Gossec, L. "Predictive factors of total hip replacement due to primary osteoarthritis: a prospective 2 year study of 505 patients". In: *Annals of the Rheumatic Diseases* 64.7 (July 2005), pp. 1028–1032. ISSN: 0003-4967. DOI: 10.1136/ard.2004.029546. URL: https://ard.bmj.com/lookup/doi/10.1136/ard.2004.029546.

[242] Roemer, F. W., Kwoh, C. K., Hannon, M. J., Hunter, D. J., Eckstein, F., Wang, Z., Boudreau, R. M., John, M. R., Nevitt, M. C., and Guermazi, A. "Can structural joint damage measured with MR imaging be used to predict knee replacement in the following year?" In: *Radiology* 274.3 (Mar. 2015), pp. 810–20. ISSN: 1527-1315. DOI: 10.1148/radiol.14140991. URL: http://www.ncbi.nlm.nih.gov/pubmed/25279436%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4455669.

[243] Hafezi-Nejad, N., Zikria, B., Eng, J., Carrino, J. A., and Demehri, S. "Predictive value of semi-quantitative MRI-based scoring systems for future knee replacement: data from the osteoarthritis initiative". In: *Skeletal Radiology* 44.11 (Nov. 2015), pp. 1655–1662. ISSN: 0364-2348. DOI: 10.1007/s00256-015-2217-2. URL: http://link.springer.com/10.1007/s00256-015-2217-2.

[244] Everhart, J. S., Abouljoud, M. M., Kirven, J. C., and Flanigan, D. C. "Full-Thickness Cartilage Defects Are Important Independent Predictive Factors for Progression to Total Knee Arthroplasty in Older Adults with Minimal to Moderate Osteoarthritis". In: *Journal of Bone and Joint Surgery* 101.1 (Jan. 2019), pp. 56–63. ISSN: 0021-9355. DOI: 10.2106/JBJS.17.01657. URL: https://journals.lww.com/00004623-201901020-00007.

[245] Tan, J., Zhang, B., Cho, K., Chang, G., and Deniz, C. M. "Semi-supervised learning for predicting total knee replacement with unsupervised data augmentation". In: *Medical Imaging 2020: Computer-Aided Diagnosis*. Ed. by Hahn, H. K. and Mazurowski, M. A. SPIE, Mar. 2020, p. 24. ISBN: 9781510633957. DOI: 10.1117/12.2551357. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/2551357/Semi-supervised-learning-for-predicting-total-knee-replacement-with-unsupervised/10.1117/12.2551357.full.

[246] Roth, M., Emmanuel, K., Wirth, W., Kwoh, C. K., Hunter, D. J., Hannon, M. J., and Eckstein, F. "Changes in Medial Meniscal Three-Dimensional Position and Morphology As Predictors of Knee Replacement in Rapidly Progressing Knee Osteoarthritis: Data From the Osteoarthritis Initiative". In: *Arthritis Care & Research* 73.7 (July 2021), pp. 1031–1037. ISSN: 2151-464X. DOI: 10.1002/acr.24193. URL: https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.24193.

[247] Jamshidi, A., Pelletier, J.-P., Labbe, A., Abram, F., Martel-Pelletier, J., and Droit, A. "Machine Learning–Based Individualized Survival Prediction Model for Total Knee Replacement in Osteoarthritis: Data From the Osteoarthritis Initiative". In: *Arthritis Care & Research* 73.10 (Oct. 2021), pp. 1518–1527. ISSN: 2151-464X. DOI: 10.1002/acr.24601. URL: https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr.24601.

[248] Almhdie-Imjabbar, A., Toumi, H., Harrar, K., Pinti, A., and Lespessailles, E. "Subchondral tibial bone texture of conventional X-rays predicts total knee arthroplasty". In: *Scientific Reports* 12.1 (May 2022), p. 8327. ISSN: 2045-2322. DOI: 10.1038/s41598-022-12083-x. URL: https://www.nature.com/articles/s41598-022-12083-x.

[249] Sindhu, N., Mishra, S., Gowrishankar, S., Anushka, S., Snehananda, H., and Veena, A. "Prediction of Knee-Replacement using Deep-Learning Approach". In: *2022 International Conference on Edge Computing and Applications (ICECAA)*. IEEE, Oct. 2022, pp. 1213–1218. ISBN: 978-1-6654-8232-5. DOI: 10.1109/ICECAA55415.2022.9936520. URL: https://ieeexplore.ieee.org/document/9936520/.

[250] Fang, J., Fan, C., and Zeng, J. "Predictive Value Analysis of MR Imaging Features on the Risk of Knee Replacement in Patients With Knee Arthritis". In: *Journal of Mechanics in Medicine and Biology* 22.08 (Oct. 2022). ISSN: 0219-5194. DOI: 10.1142/S0219519422400309. URL: https://www.worldscientific.com/doi/10.1142/S0219519422400309.

[251] Mahmoud, K., Alagha, M. A., Nowinka, Z., and Jones, G. "Predicting total knee replacement at 2 and 5 years in osteoarthritis patients using machine learning". In: *BMJ Surgery, Interventions, & Health Technologies* 5.1 (Feb. 2023), e000141. ISSN: 2631-4940. DOI: 10.1136/bmjsit-2022-000141. URL: https://sit.bmj.com/lookup/doi/10.1136/bmjsit-2022-000141.

[252] Rajamohan, H. R., Wang, T., Leung, K., Chang, G., Cho, K., Kijowski, R., and Deniz, C. M. "Prediction of total knee replacement using deep learning analysis of knee MRI". In: *Scientific Reports* 13.1 (2023).

[253] Yang, L., Xiao, F., and Cheng, C. "Knee Replacement Risk Prediction Modeling for Knee Osteoarthritis Using Clinical and Magnetic Resonance Image Features: Data From The Osteoarthritis Initiative". In: *Journal of Mechanics in Medicine and Biology* 23.08 (Oct. 2023). ISSN: 0219-5194. DOI: 10.1142/S0219519423400687. URL: https://www.worldscientific.com/doi/10.1142/S0219519423400687.

[254] Driban, J. B., Lu, B., Flechsenhar, K., Lo, G. H., and McAlindon, T. E. "The Prognostic Potential of End-Stage Knee Osteoarthritis and Its Components to Predict Knee Replacement: Data From the Osteoarthritis Initiative". In: *The Journal of Rheumatology* 50.11 (Nov. 2023), pp. 1481–1487. ISSN: 0315-162X. DOI: 10.3899/jrheum.2023-0017. URL: http://www.jrheum.org/lookup/doi/10.3899/jrheum.2023-0017.

[255] Saxer, F., Demanse, D., Brett, A., Laurent, D., Mindeholm, L., Conaghan, P., and Schieker, M. "Prognostic value of B-score for predicting joint replacement in the context of osteoarthritis phenotypes: Data from the osteoarthritis initiative". In: *Osteoarthritis and Cartilage Open* 6.2 (June 2024), p. 100458. ISSN: 26659131. DOI: 10.1016/j.ocarto.2024.100458. URL: https://linkinghub.elsevier.com/retrieve/pii/S2665913124000256.

[256] Zhang, J., Jiang, T., Chan, L.-C., Lau, S.-H., Wang, W., Teng, X., Chan, P.-K., Cai, J., and Wen, C. "Radiomics analysis of patellofemoral joint improves knee replacement risk prediction: Data from the Multicenter Osteoarthritis Study (MOST)". In: *Osteoarthritis and Cartilage* 6.2 (2024).

[257] Wang, W. and Kuang, S. "Role of the WOMAC Scores in Preoperative Decision-Making and Analysis of Knee Replacement for Knee Osteoarthritis Patients". In: *Journal of Mechanics in Medicine and Biology* 23.08 (Oct. 2023). ISSN: 0219-5194. DOI: 10.1142/S0219519423400705. URL: https://www.worldscientific.com/doi/10.1142/S0219519423400705.

[258] Heisinger, S., Hitzl, W., Hobusch, G. M., Windhager, R., and Cotofana, S. "Predicting Total Knee Replacement from Symptomology and Radiographic Structural Change Using Artificial Neural Networks—Data from the Osteoarthritis Initiative (OAI)". In: *Journal of Clinical Medicine* 9.5 (May 2020), p. 1298. ISSN: 2077-0383. DOI: 10.3390/jcm9051298. URL: https://www.mdpi.com/2077-0383/9/5/1298.

[259] Yang, D., Sinha, T., Adamson, D., and Rose, C. ""Turn on, Tune in, Drop out": Anticipating student dropouts in Massive Open Online Courses". In: *Proceedings of the NIPS Workshop on Data Driven Education* (2013), pp. 1–8.

[260] Xu, Y., Xiong, H., Liu, W., Liu, H., Guo, J., Wang, W., Ruan, H., Sun, Z., and Fan, C. "Development and Validation of a Deep-Learning Model to Predict Total Hip Replacement on Radiographs". In: *Journal of Bone and Joint Surgery* 106.5 (Mar. 2024), pp. 389–396. ISSN: 0021-9355. DOI: 10.2106/JBJS.23.00549. URL: https://journals.lww.com/10.2106/JBJS.23.00549.

[261] Agricola, R., Reijman, M., Bierma-Zeinstra, S., Verhaar, J., Weinans, H., and Waarsing, J. "Total hip replacement but not clinical osteoarthritis can be predicted by the shape of the hip: a prospective cohort study (CHECK)". In: *Osteoarthritis and Cartilage* 21.4 (Apr. 2013), pp. 559–564. ISSN: 10634584. DOI: 10.1016/j.joca.2013.01.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S1063458413000198.

[262] Qiu, R., Jia, Y., Wang, F., Divakarmurthy, P., Vinod, S., Sabir, B., and Hadzikadic, M. "Predictive Modeling of the Total Joint Replacement Surgery Risk: a Deep Learning

Based Approach with Claims Data." In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2019 (2019), pp. 562–571. ISSN: 2153-4063. URL: http://www.ncbi.nlm.nih.gov/pubmed/31259011%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6568108.

[263] Li, H., Chan, L., Chan, P., and Wen, C. "An interpretable knee replacement risk assessment system for osteoarthritis patients". In: *Osteoarthritis and Cartilage Open* 6.2 (June 2024), p. 100440. ISSN: 26659131. DOI: 10.1016/j.ocarto.2024.100440. URL: https://linkinghub.elsevier.com/retrieve/pii/S2665913124000074.

[264] Birrell, F., Afzal, C., Nahit, E., Lunt, M., Macfarlane, G. J., Cooper, C., Croft, P. R., Hosie, G., and Silman, A. J. "Predictors of hip joint replacement in new attenders in primary care with hip pain." In: *The British journal of general practice : the journal of the Royal College of General Practitioners* 53.486 (Jan. 2003), pp. 26–30. ISSN: 0960-1643. URL: http://www.ncbi.nlm.nih.gov/pubmed/12564273%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1314488.

[265] McLaughlin, J., Kipping, R., McLeod, H., Judge, A., and Owen-Smith, A. "Health optimisation for patients with obesity before elective orthopaedic surgery: a qualitative study of professionals' views on restrictive approaches and future practice". In: *Perioperative Medicine* 13.1 (2024), p. 104.

[266] Cai, H., Que, Z., Chen, J., Chen, D., Rui, G., and Lan, W. "Association between different insulin resistance surrogates and osteoarthritis: a cross-sectional study from NHANES 1999–2018". In: *BMC Musculoskeletal Disorders* 25.1 (2024), p. 901.

[267] Madsen, H. J., Gillette, R. A., Colborn, K. L., Henderson, W. G., Dyas, A. R., Bronsert, M. R., Lambert-Kerzner, A., and Meguid, R. A. "The association between obesity and postoperative outcomes in a broad surgical population: a 7-year American College of Surgeons National Surgical Quality Improvement analysis". In: *Surgery* 173.5 (2023), pp. 1213–1219.

[268] Akinrinmade, A. O., Adebile, T. M., Ezuma-Ebong, C., Bolaji, K., Ajufo, A., Adigun, A. O., Mohammad, M., Dike, J. C., and Okobi, O. E. "Artificial Intelligence in Healthcare: Perception and Reality." In: *Cureus* 15.9 (Sept. 2023), e45594. ISSN: 2168-8184. DOI:

10.7759/cureus.45594. URL: http://www.ncbi.nlm.nih.gov/pubmed/37868407%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC10587915.

[269] Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., and Liu, N. "Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies". In: *Journal of Biomedical Informatics* 126 (Feb. 2022), p. 103980. ISSN: 15320464. DOI: 10.1016/j.jbi.2021.103980. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046421003099.

[270] Zhao, Z., Shi, Y., Wu, S., Yang, F., Song, W., and Liu, N. "Interpretation of Time-Series Deep Models: A Survey". In: (May 2023). URL: http://arxiv.org/abs/2305.14582.

[271] Huang, Q., Yamada, M., Tian, Y., Singh, D., and Chang, Y. "GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks". In: *IEEE Transactions on Knowledge and Data Engineering* 35.7 (July 2023), pp. 6968–6972. ISSN: 1041-4347. DOI: 10.1109/TKDE.2022.3187455. URL: https://ieeexplore.ieee.org/document/9811416/.

[272] Kasanishi, T., Wang, X., and Yamasaki, T. "Edge-Level Explanations for Graph Neural Networks by Extending Explainability Methods for Convolutional Neural Networks". In: *2021 IEEE International Symposium on Multimedia (ISM)*. IEEE, Nov. 2021, pp. 249–252. ISBN: 978-1-6654-3734-9. DOI: 10.1109/ISM52913.2021.00049. URL: https://ieeexplore.ieee.org/document/9666067/.

[273] Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., and Medya, S. "A Survey on Explainability of Graph Neural Networks". In: (June 2023). URL: http://arxiv.org/abs/2306.01958.

[274] Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. "GNNExplainer: Generating explanations for graph neural networks". In: *Advances in Neural Information Processing Systems* 32.iii (2019). ISSN: 10495258.

[275] Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. "Explainability Methods for Graph Convolutional Neural Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 10764–10773.

ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.01103. URL: https://ieeexplore.ieee.org/document/8954227/.

[276] Leavitt, M. L. and Morcos, A. "Towards falsifiable interpretability research". In: (Oct. 2020). URL: http://arxiv.org/abs/2010.12016.

[277] Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., and Pardalos, P. M. "A review of Explainable Artificial Intelligence in healthcare". In: *Computers and Electrical Engineering* 118 (Aug. 2024), p. 109370. ISSN: 00457906. DOI: 10.1016/j.compeleceng.2024.109370. URL: https://linkinghub.elsevier.com/retrieve/pii/S0045790624002982.

[278] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74. URL: http://ieeexplore.ieee.org/document/8237336/.

[279] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. "The (Un)reliability of Saliency Methods". In: 2019, pp. 267–280. DOI: 10.1007/978-3-030-28954-6{_}14. URL: http://link.springer.com/10.1007/978-3-030-28954-6_14.

[280] Su, C., Gao, S., and Li, S. "GATE: Graph-Attention Augmented Temporal Neural Network for Medication Recommendation". In: *IEEE Access* 8 (2020), pp. 125447–125458. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3007835. URL: https://ieeexplore.ieee.org/document/9134772/.

[281] Sun, C., Dui, H., and Li, H. "Interpretable time-aware and co-occurrence-aware network for medical prediction". In: *BMC Medical Informatics and Decision Making* 21.1 (Dec. 2021), p. 305. ISSN: 1472-6947. DOI: 10.1186/s12911-021-01662-z. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01662-z.

[282] Sun, Z., Dong, W., Shi, J., He, K., and Huang, Z. "Attention-Based Deep Recurrent Model for Survival Prediction". In: *ACM Transactions on Computing for Healthcare* 2.4 (Oct. 2021), pp. 1–18. ISSN: 2691-1957. DOI: 10.1145/3466782. URL: https://dl.acm.org/doi/10.1145/3466782.

[283] Yue, C., Cui, G., Ma, M., Tang, Y., Li, H., Liu, Y., and Zhang, X. "Associations between smoking and clinical outcomes after total hip and knee arthroplasty: A systematic review and meta-analysis". In: *Frontiers in Surgery* 9 (2022), p. 970537.

[284] Powell, J. and Buchan, I. "Electronic Health Records Should Support Clinical Research". In: *Journal of Medical Internet Research* 7.1 (Mar. 2005), e4. ISSN: 1438-8871. DOI: 10.2196/jmir.7.1.e4. URL: http://www.jmir.org/2005/1/e4/.

[285] Sohal, K., Mason, D., Birkinshaw, J., West, J., McEachan, R. R., Elshehaly, M., Cooper, D., Shore, R., McCooe, M., Lawton, T., Mon-Williams, M., Sheldon, T., Bates, C., Wood, M., and Wright, J. "Connected Bradford: a Whole System Data Linkage Accelerator". In: *Wellcome Open Research* 7 (Nov. 2022), p. 26. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.17526.2. URL: https://wellcomeopenresearch.org/articles/7-26/v2.

[286] Chen, J., Yin, C., Wang, Y., and Zhang, P. "Predictive modeling with temporal graphical representation on electronic health records". In: *IJCAI: proceedings of the conference*. Vol. 2024. 2024, p. 5763.

[287] Chen, F., Wang, L., Hong, J., Jiang, J., and Zhou, L. "Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models". In: *Journal of the American Medical Informatics Association* 31.5 (2024), pp. 1172–1183.

[288] Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., and Hassanpour, S. "Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice". In: *PLOS digital health* 4.4 (2025), e0000810.

[289] Kontopantelis, E., Stevens, R. J., Helms, P. J., Edwards, D., Doran, T., and Ashcroft, D. M. "Spatial distribution of clinical computer systems in primary care in England

in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study". In: *BMJ open* 8.2 (2018), e020738.

[290] Digital, N. *Clinical Practice Research Datalink - NDRS*. 2025. URL: https://digital.nhs.uk/ndrs/our-work/ncras-partnerships/cprd.

[291] Chaudhry, Z., Mannan, F., Gibson-White, A., Syed, U., Ahmed, S., Kousoulis, A., and Majeed, A. "Outputs and growth of primary care databases in the United Kingdom: bibliometric analysis." In: *Journal of innovation in health informatics* 24.3 (2017), p. 942.

Appendix A

Systematic Literature Review Appendix

A.1 PRISMA Checklist

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Page 1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Page 1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Page 2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Page 1-2
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Page 2
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Page 2
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Page 2
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Appendix
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Page 3
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Page 3 & Appendix
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Appendix
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Page 3
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	N/A

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Page 3-5
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Page 3 & Appendix
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Page 6-10
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Page 4
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	N/A
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/A
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Page 3-5
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	N/A
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Page 12
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Page 13
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Page 13

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

A.2 Search Strings

Below we provide details of the search terms for each database including how many papers were retrieved from each.

The search terms were combined using Boolean operators (OR, AND) and adjacency keys to create database queries. The search term query generated was as follows: ((graph OR graphs OR graph-based OR (node* AND edge*) OR “knowledge graph” OR “network analysis”) AND (”electronic health record” OR “medical records systems” OR “record-linkage” OR (routine adj5 data) OR ((electronic OR link* OR compute* OR anonymi?ed) adj5 record) OR ((health OR patient OR clinic* OR medic* OR case) adj5 (record* OR data OR plan* OR chart*) adj5 (compute* OR system OR electronic OR link* OR dataset OR network)) OR EMR OR EPR OR EHR) AND (predict* OR diagnos* OR prognos*))

Search conducted on: 27/02/2023

MEDLINE. Papers retrieved: 303 ((graph or graphs or graph-based or (node* and edge*) or “knowledge graph” or “network analysis”) and (”electronic health record” or “medical records systems” or “record-linkage” or (routine adj5 data) or ((electronic or link* or compute* or anonymi?ed) adj5 record) or ((health or patient or clinic* or medic* or case) adj5 (record* or data or plan* or chart*) adj5 (compute* or system or electronic or link* or dataset or network)) or EMR or EPR or EHR) and (predict* or diagnos* or prognos*))

Web of Science. Papers retrieved: 410 AB=(((graph OR graphs OR graph-based OR (node* AND edge*) OR “knowledge graph” OR “network analysis”) AND (”electronic health record” OR “medical records systems” OR “record-linkage” OR (routine near/5 data) OR ((electronic OR link* OR compute* OR anonymi?ed) near/5 record) OR ((health OR patient OR clinic* OR medic* OR case) near/5 (record* OR data OR plan* OR chart*) near/5 (compute* OR system OR electronic OR link* OR dataset OR network))OR EMR OR EPR OR EHR) AND (predict* OR diagnos* OR prognos*)))

Scopus. Papers retrieved: 633 TITLE-ABS-KEY (((graph OR graphs OR graph-based OR (node* AND edge*) OR “knowledge graph” OR “network analysis”) AND (“electronic health record” OR “medical records systems” OR “record-linkage” OR (routin* W/5 data) OR ((electronic OR link* OR compute* OR anonymi?ed) W/5 record) OR ((health OR patient OR clinic* OR medic* OR case) W/5 (record* OR data OR plan* OR chart*) W/5 (compute*

OR system OR electronic OR link* OR dataset OR network)) OR emr OR epr OR ehr) AND (predict* OR diagnos* OR prognos*)))

Total papers (including potential duplicates) = 1,346

A.3 Data Extraction Items

The bullet list below shows all of the items we extracted from the papers during the data extraction process of our systematic literature review.

1. First author: Surname of first author.
2. Publication date: Year the study was published.
3. DOI number: DOI number of the study (this could also be arVix number if no DOI available).
4. Title: Title of paper.
5. Journal/ conference title: Journal or conference the paper was submitted to.
6. Model description/method: What method or ML model was used e.g. CNN, LSTM with a description.
7. Method for selection of predictors for inclusion in modelling: e.g. PCA, regression.
8. Patient inclusion criteria: What inclusion/exclusion criteria did the authors use.
9. Missing data: Type of data missing, handling of missing data.
10. Disease/ condition/ health setting: Health domain e.g. Cancer, MSK. Start point.
11. Dataset/ participant description: e.g. Snomed codes from over 40 year olds from West Yorkshire UK. Patient population. Where data is from geographically.
12. Sample size: Total number of peoples data used in the model and number of outcomes/ events.
13. Label type/ health outcome(s) to be predicted: Health outcome such as frailty. Endpoint.
14. Prognostic or diagnostic outcome?: State whether outcome is a prognostic prediction or a diagnostic prediction.

15. Predictors in final model: e.g. age, sex, codes.
16. Node allocation: What were the nodes used for in the graph.
17. Edge allocation: What where the edges used for in the graph.
18. Timing of predictor measurement: e.g. at patient presentation, at diagnosis, at treatment initiation.
19. Node, edge or graph-level prediction?: What type of graph prediction did the authors use.
20. Predictive performance (AUROC)
21. Predictive performance (AUPRC)
22. Predictive performance (Accuracy)
23. Predictive performance (Precision)
24. Predictive performance (Recall)
25. Predictive performance (Specificity)
26. Predictive performance (F1)
27. True Positive (TP)
28. True Negative (TN)
29. False Positive (FP)
30. False Negative (FN)
31. Predictive performance (. . .): Other predictive performance metrics not included.
32. External validation(s): Have the authors externally validated their model (using a different dataset with external researchers).
33. Internal validation(s) : Has the model been internally validated?
34. Comparison studies on the same dataset: Techniques and predictive performance of these comparison studies. Ablation, comparison models from the literature and baselines.
35. Interpretation of presented models: Are the models ready for real-world use or more research required?

36. Are clinicians involved in the study?: Clinical author or involvement via contribution.

37. Miscellaneous: Any other information from the paper that seems interesting.

A.4 Screening

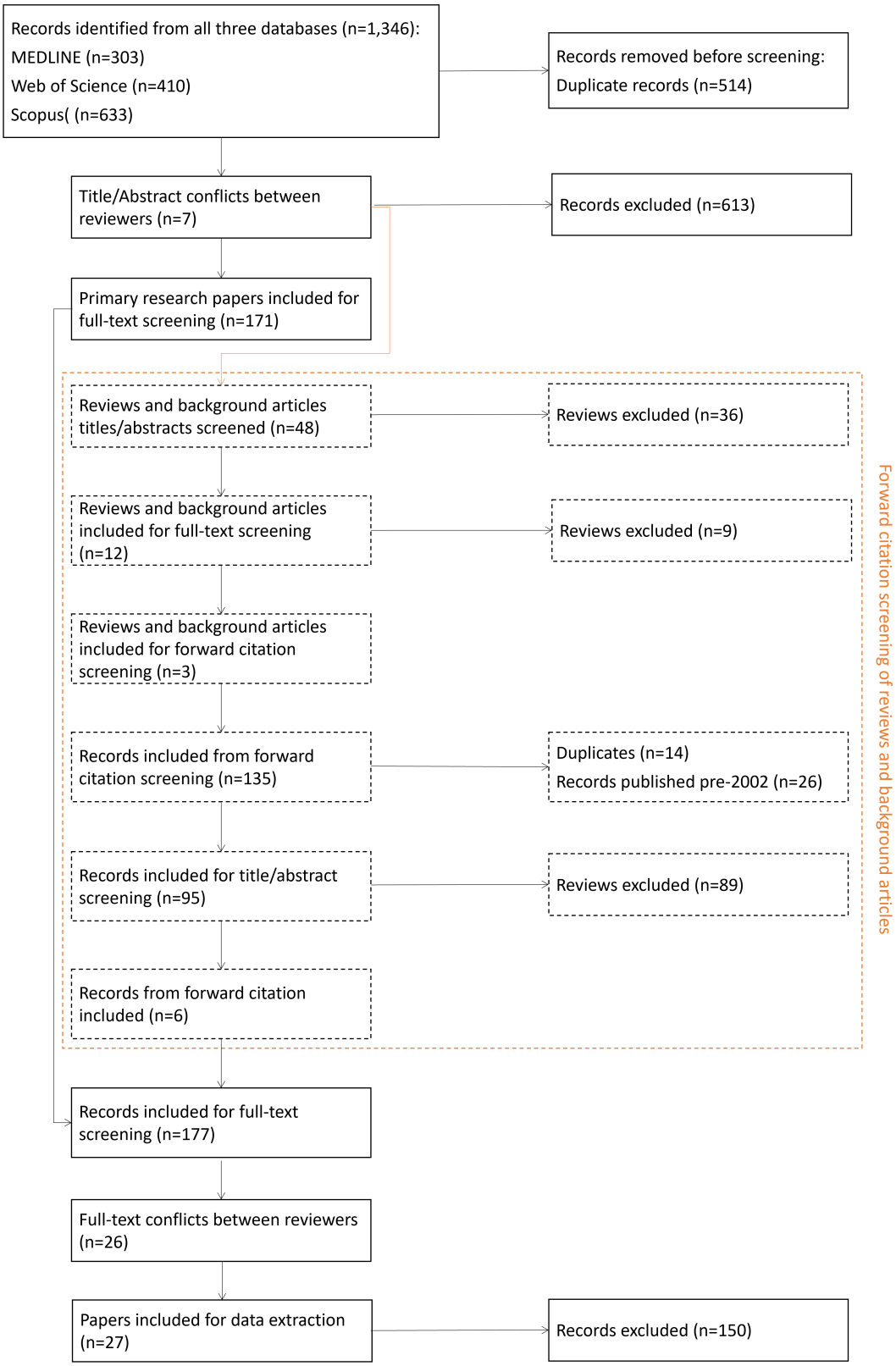


Figure A.1: Full screening process flowchart.

A.5 Study Characteristics

Table A.1: Overview of included studies (1/9)

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[139]	Hospital EHR system (Australia)	1,800	Admission information, diagnoses, procedures (ICD-10, diagnosis related group, Australian National Subacute and Non-Acute Patient Classification (AN-SNAP))	-	Risk of DM	1. Social Network analysis of graph 2. Longitudinal Distance Matching	None	-
[140]	Private healthcare hospital data (Australia)	38,500	ICD-10 codes and administrative data	-	Risk of DM	1. LR 2. Decision tree	Internal Val 6:3:1	Accuracy
[158]	Medical system (City in North China)	7,372	ICD-10 codes	-	Top K Diseases	**Collaborative assessment prediction model (CAPM)** 1. Temporal graph representation of events 2. Vector similarity of temporal graph representation	Leave one out	Precision
[142]	NHIRD (Taiwan)	1,000,000	ICD9-CM Anatomical Therapeutic Chemical Classification (ATC) codes	10 Years	Success of pharmacological prevention of complications of: a. HTN b. Hyperlipidaemia c. DM	**Cross-Global Attention Graph Kernel Network** 1. EHR Graph embedding from GCN 2. Global node attention/cluster learning 3. SVM classifier based on cosine distance of pairs	Internal Val 8:1:1	AUROC, Accuracy, F1 Score

Table A.2: Overview of included studies (2/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[141]	MIMIC-III (USA)	5,956	ICD9-CM, Laboratory Investigations	6, 12, 24, 48 hours	Mortality	**Heterogenous graph embedding** 1. CNN to build an embedding representation of ICD9 and Laboratory investigations 2. SoftMax classifier	10-fold CV	AUROC, AUPRC
[159]	MIMIC-III (USA)	34,560	a. Clinical Discharge Note b. Clinical notes from the first three days of admission	30 days	30-day hospital readmission	**DeepNote-GNN** 1. Clinical-BERT develop embedding vector of clinical discharge 2. Admission feature vector from admission 3. GCN construct a node feature vector 4. Classiify on cosine similarity	5-fold CV	AUROC, AUPRC, Recall @ 80% Precision
[146]	NHIRD (Taiwan)	1,000,000	ICD9-CM ATC codes	One month	Antibiotic Failure / Success	**Graph Kernel** 1. Temporal graph construction of EHR 2. SVM classifier	10-fold CV	AUROC, Accuracy, F1 Score
[147]	Unknown	34,427	ICD9-CM, medication, laboratory investigations	90 days	Risk prediction of pairs of: a. HF b. DM c. Chronic Kidney Disease (CKD) d. COPD	**Heterogeneous CNN** 1. CNN on a graph representation of temporal events of EHRs to learn vector representation 2. MLP/SoftMax classifier	-	Precision, Recall, F1 Score

Table A.3: Overview of included studies (3/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[160]	MIMIC-III (USA)	7,537	ICD9-CM	-	a. Mortality b. Readmission c. Sepsis / HF Prediction d. Next Diagnosis	Time-aware & **Co-Occurrence aware Network (TCoN)** 1. Time-GRU to develop a representation of admission/visit. 2. Two-headed attention to GRU embeddings 3. SoftMax classifier	5-fold CV	AUROC, AUPRC, Accuracy @ k-value values (5,15,25,35)
[161]	National Cancer Registry (Colorectal Cancer) (UK)	54,000	ICD10-CM	-	Next Diagnosis	**Deep Diffusion Process** 1. LSTM on the temporal sequence of EHRs to model the next event 2. Hidden state provides transmission function of a point process model	-	AUROC
[145]	eICU (USA)	41,026	ICD9-CM, CPT codes, Laboratory Investigations	-	a. Readmission b. Mortality in episode c. Next Diagnosis	**Graph Convolutional Transformer (GCT)** 1. GCT to learn embedding from the graph representation of visits and co-occurrence 2. Classification on embedding	5-fold CV	AUPRC Accuracy

Table A.4: Overview of included studies (4/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[68]	Foundation Medicine Inc & Mayo Clinic EHR (USA)	1,011	Oncology genetic reports, Medication, Laboratory Investigations, ICD9-CM, family history.	0 - 24 Months	Cancer type classification	**FHIR2RDF** 1. Graph representation using FHIR2RDF and Node2vec to extract features 2. ML methods for downstream (RF, Naïve Bayes, LR, SVM, CNN, MLP, GCN)	10-fold CV	AUROC
[162]	Proprietary EHR Dataset (Unknown)	Unknown	10,713 for HF cohort 3,830 for chronic liver disease cohort & Patient Demographics, Hospital utilisation information, ICD10-CM, Procedural Codes	30 days	a. Readmission b. Mortality during admission	**MedGraph** 1. Bipartite graph construction of code occurrence and visit. 2. Structural learning using Wasserstein distance 3. Temporal learning using LSTM 4. SoftMax Classification	75:25 Train/Test Split	AUROC, AUPRC
[163]	1. Geriatric health examination cohort study 2. National Death Registry (Taiwan)	102,258	Demographics, habits, laboratory investigations, physical examinations, mental health assessments, cognitive function	Three years	a. Mortality Risk b. Diagnosis of death	**Semi-supervised heterogeneous graph (SHG)-Health** 1. Graph construction of heterogeneous data 2. Semi-supervised learning on unlabelled data 3. Classification	5-fold CV	Precision, Recall, F1-Score

Table A.5: Overview of included studies (5/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[148]	Proprietary EHR dataset (North China)	92,652	ICD10-CM, Time stamps, Demographics, Clinical event	180 days	Diagnosis of: a. COPD b. IHD	**MTPGraph** 1. Patient temporal profiling 2. Temporal feature extraction/feature map construction 3. SVM classification on the feature vector	10-fold CV	AUROC, Recall, Accuracy
[164]	NHIRD (Taiwan)	1,000,000	Demographics, ICD9-CM ATC for drug prescriptions, Time stamps	14 days	Success of Antibiotic: a. Pneumonia b. AoM c. Acute cystitis d. UTI	**Multiple Graph Kernel Fusion (MGKF)** 1. Temporal signature using graph kernel 2. Fusion of pairwise temporal proximity and kernel embedding using MLP 3. Classification using MLP + Sigmoid output	10-fold CV	AUROC, Accuracy, F1-Score
[165]	MIMIC-II (USA)	1,170	ICD9-CM, Laboratory Investigations, Hospital admission information, demographics	30 Days	Readmission within 30 days	**Subgraph Augmented Non-negative Matrix Factorization (SANMF)** 1. SANMF convert time-series data into graphs 2. Unsupervised learning clustering of sub-graph trends to extract features 3. LR classification using demographics and subgraph features	5-fold CV	AUROC, Accuracy, Precision, Recall, Negative Predictive Value (NPV)

Table A.6: Overview of included studies (6/9)

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[166]	MIMIC-III (USA)	48,393	Discharge summary Entity from UMLS	30 Days	Unplanned readmission within 30 days	1. Graph construction of UMLS Entity Discharge summary 2. GCN and attention layer encoding of entity relationship 3. Bi-LSTM encoding of a document 4. MLP Classifier with GCN/Bi-LSTM embedding	Internal Val 8:1:1	AUROC, AUPRC, Recall @ 80% Precision
[167]	1. eICU 2. Paediatric EHRs in 'Tertiary centre' (USA, China)	1. 41,026 2. 144,170	1. ICD9-CM, Medication, Procedure codes, Medical Examination 2. Symptoms, medical examination, disease, drugs	-	a. Readmission to ICU b. Mortality during ICU admission c. URTI	**Statistics and Knowledge-based Graph Transformer (S_K_GT)** 1. Graph representation of the clinical episode 2. Transformer to form conditional probability matrix & Attention matrix 3. Fine-tune for the prediction task	Internal Val 8:1:1	AUROC
[168]	Proprietary EHREHR dataset from seven hospitals (Australia)	Cancer: 4,293 Acute Myocardial Infarction (AMI): 2,941	ICD10-CM, Procedure codes, demographic	One year 30 days	a. One-year Mortality b. 30 days Readmission	**Graph SVM** 1. Graph generation frherEHR using Jaccard Index 2. Alternating direction method of multipliers (ADMM) to train SVM Classifier	100 Iterations Mean values presented	AUROC, PPV, Sensitivity, Specificity, F1-score

Table A.7: Overview of included studies (7/9)

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method
[169]	MIMIC-III (USA)	3,431,622	Demographics, Vital signs	24 hours	Death/ Decompensation	**Temporal Aware Graph Convolution Network (TAGNet)** 1. Graph generation on of EHR 2. GRU encoding at each timestep 3. GCN learns representation from attention from GRU hidden states 4. MLP Classifier	Internal Va 70:15:15
[118]	Real-world EHR data warehouse (Unknown)	319,650	ICD9-CM	One year	a. One-year hospitalisation b. Onset of HF	**Temporal Phenotyping** 1. Temporal graph construction on the seqsequence of EHR events 2. Temporal phenotyping using four stated methods in vector representation 3. SVM classifier	10-fold CV

. CV = Cross-fold validation.

Table A.8: Overview of included studies (8/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[170]	NHIRD (Taiwan)	UTI: 1,501,310 AoM: 151,522 Community Acquired Pneumonia (CAP): 95,796 Cystitis: 733,119 HTN: 235,695 Hyperlipaemia: 123,380 DM: 131,997	Demographics, ICD9-CM ATC Code	UTI / AoM, CAP, Cys: 14 Days HTN, DM Hyperlipaemia: 5 years	Failure of treatment 1. Patient Graph Construction 2. Graph Kernel pairwise matching 3. MGKF [164] / Cross-global [142] frameworks used to classify outcome	10-fold CV	AUROC, Accuracy, F1-Score	
[171]	EHR from In Vitro Fertilisation (IVF) clinic (South Korea)	269	Demographics, IVF Treatment variables	-	Positive Pregnancy Test	**Bayesian Networks** 1. Joint probability of conditional dependencies of selected features 2. Classification using conditional probabilities	10-fold CV	Accuracy

Table A.9: Overview of included studies (9/9). CV = Cross-fold validation.

Ref	Dataset (Location)	Sample Size	Description of Data Extracted	Prediction Time Horizon	Outcome(s)	ML Technique	Validation Method	Metrics Presented
[172]	a. NYU YU Langone Health EHR b. MIMIC-III c. eICU (USA)	a.1,600,000 b. 50,000 c. 41,000	a. ICD10-CM, Logical Observation Identifiers Names and Codes (LOINC), medication b-c. ICD9-CM, Laboratory Investigations (bucketed), CPT	a. 12-24 months b-c. -	a. Alzheimer's Disease Onset b. Mortality during admission c. Readmission	**Variationally Regularized Graph Repreeesentation (VRFR)** 1. EHR features encoded into embeddings 2. Multi-head attention to learn encoded graph 3. Decoder classifier with latent space regularisation	Internal Val 8:1:1	AUPRC, PPV @ 40% Recall
[173]	CardioNet (South Korea)	53,841	EHR Data from Seoul Asan Medical Center, diagnosis, laboratory investigations, echocardiology, physical, medication, surgery, smoking	Observed for 5 years after being admitted with angina	Cardiovascular disease: whether mortality, MI, stroke or HF occurred in the 5-year period	**Heterogeneous bipartite graphs** Node embeddings aggregated and the target node represented. Neighbouring nodes encoded into embeddings and updated.	Train and validation split not specified	AUROC, AUPRC
[174]	a. MIMIC-IV b. eICU (USA)	a.HF mortality: 15,528 Sepsis: 58,230 b. 150,644	a. ICD ICU population, admission, diagnosis and treatment b. ICD-9 and CPT procedure codes	a-b. First 48 hours in the ICU	a. HF mortality and sepsis b. HF mortality	**Gated Graph Attention Network** 1.Time series features and clinical events extracted from EHR. 2. Self-attention. 3. GNN model	5-fold CV	AUROC, AUPRC, Accuracy

A.6 Node and Edge Allocation

Table A.10: Node and edge allocation types in the graphs used in the selected papers (1/2).

Papers	Node Allocation/ Use	Edge Allocation
[139]	Diagnosis	1) Number of times two diseases occurred simultaneously 2) Sequential directionality/ ordering 3) Number of times two diseases occurred sequentially (one after another)
[118]	Diagnosis, medication	Temporal proximity weighting
[167]	Diagnosis, treatment, physical examinations, symptoms	Medical relation between nodes
[141]	Diagnosis, laboratory investigations, patients	Testing or diagnosis of a patient undertaken
[145]	Diagnosis, laboratory investigations, treatment	Association weighting between nodes
[147]	Diagnosis, EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic)	Temporal proximity weighting
[148]	Diagnosis, medication, laboratory investigations	Time difference/elapsed between each node, temporal proximity weighting
[158]	Diagnosis, EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic), medication	Weights higher if two medical events are more often and closer
[68]	Diagnosis, demographics, laboratory investigations, patients, genetic data, family history	Association weighting between nodes
[162]	Diagnosis, demographics, clinical note representation, visits	Time difference/elapsed between each node, different interactions (e.g. code to timestep)
[164]	Diagnosis, demographics, medication	Time difference/elapsed between each node
[173]	Diagnosis, medication, laboratory investigations, patients, physical examinations, visits, smoking, echocardiography	Relationship between patient and medical node (e.g. edge exists between patient and smoke if the patient smokes)
[142]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic), demographics (e.g. age, BMI, gender)	Time difference/elapsed between each node
[146]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic)	Time difference/elapsed between each node, link to demographics (as the first node)
[160]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic)	Linking of nodes/EHR codes happening on the same visit
[161]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic)	Risk of disease
[170]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic), demographics (e.g. age, BMI, gender)	Time difference/elapsed between each node, link to demographics (as the first node)

Table A.11: Node and edge allocation types in the graphs used in the selected papers (2/2).

Papers	Node Allocation/ Use	Edge Allocation
[163]	Demographics (e.g. age, BMI, gender), physical examinations, mental tests, habits	Temporal proximity weighting
[169]	Demographics (e.g. age, BMI, gender), heart rate, blood pressure and oxygen saturation, eye-opening and verbal response	Fully connected initially and updated by attention
[174]	Medication, treatment, patients	Events happening on the same time step are linked via edge and weighting is value from laboratory test, or infusion drug. If patient took a prescription the edge weight is 1 otherwise it is 0 to the prescription node
[171]	Treatment, symptoms	The conditional probability of a connection between 2 nodes
[159]	Clinical note representation	The similarity between 2 nodes
[140]	Comorbidity occurrence count	Number of times two diseases occurred simultaneously, number of times two diseases occurred sequentially (one directly after another)
[166]	Average values of word embeddings from: unique words from clinical free text or the linked umls	1) Intradocument interaction level. 2) Path lengths between entity nodes. 3) String similarities based on word overlap. 4) Coine similarities
[165]	Discretized measurements of variables at a point in time	Sequential directionality/ ordering, labelling of change of quantifiable variable (up, down or no change)
[168]	EHR features	The similarity between two nodes
[172]	EHR codes (e.g. ICD-10), but which type(s) are unclear (e.g. diagnoses, demographic)	Fully connected initially and updated by attention

A.7 AUROC Baseline Model Comparison

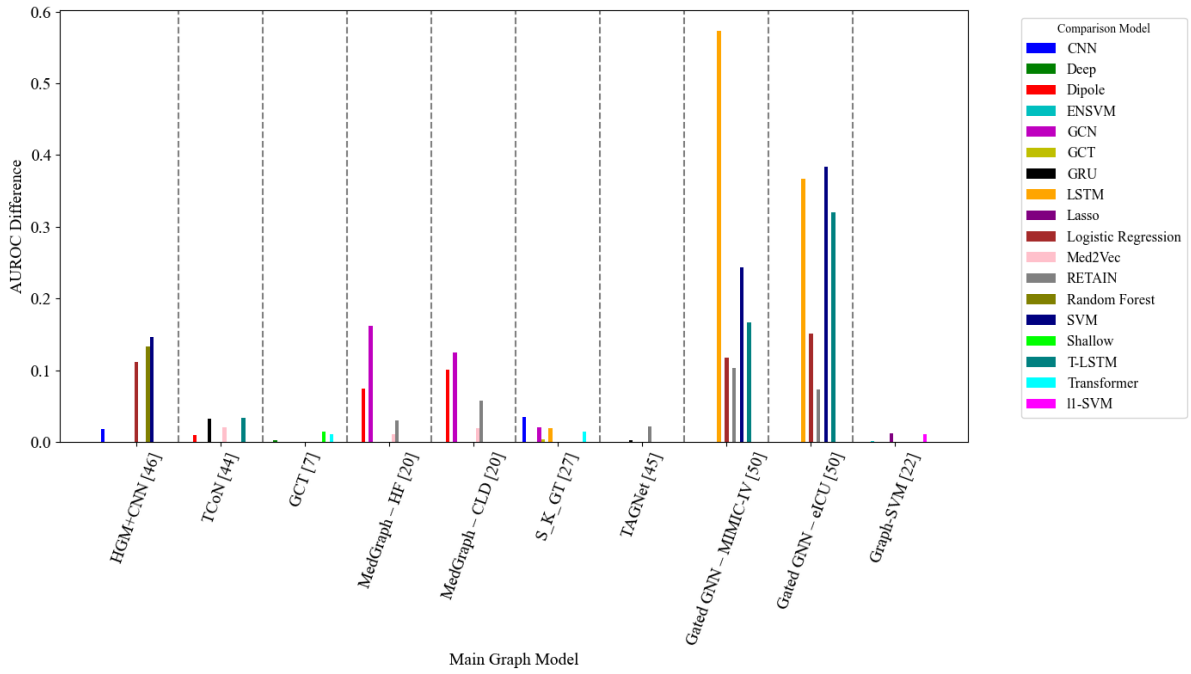


Figure A.2: Comparison of AUROC scores for mortality prediction between the primary model and alternative/baseline models presented in various studies. CL = Chronic liver disease, HF = Heart Failure.

Appendix B

Hip Replacement Prediction Appendix

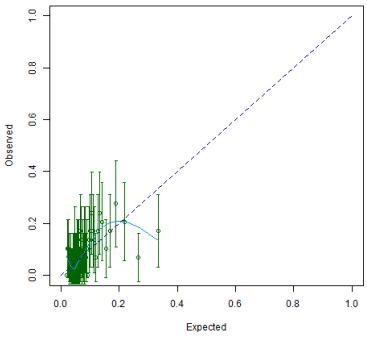


Figure B.1: Males only.

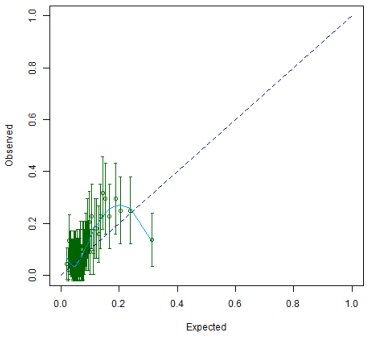


Figure B.2: Females only.

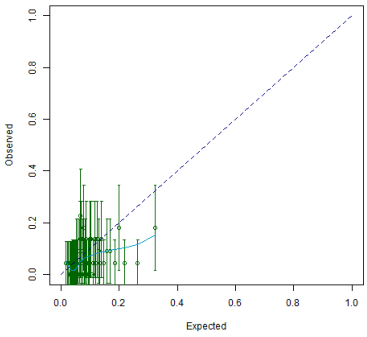


Figure B.3: 40-60 year olds.

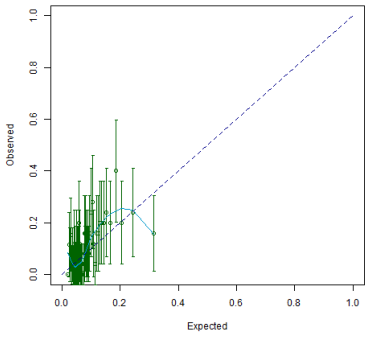


Figure B.4: 60-70 year olds.

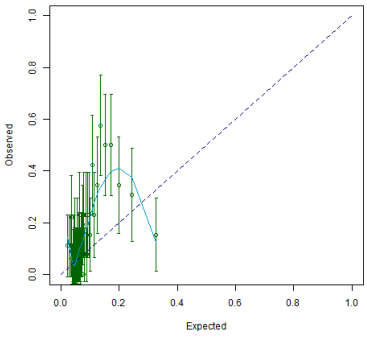


Figure B.5: 70+ year olds.

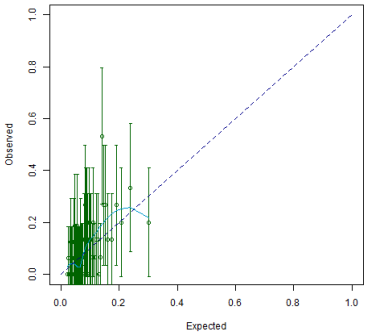


Figure B.6: IMD 1.

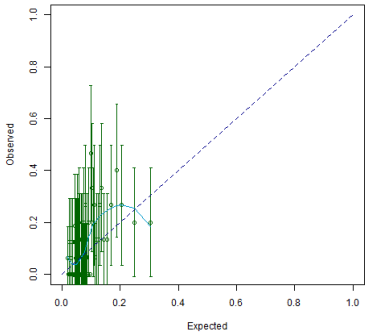


Figure B.7: IMD 2.

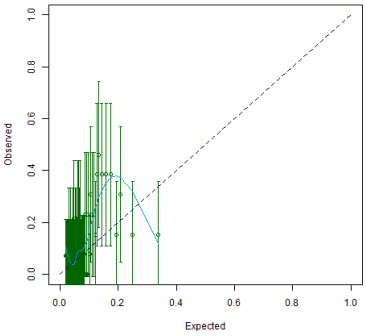


Figure B.8: IMD 3.

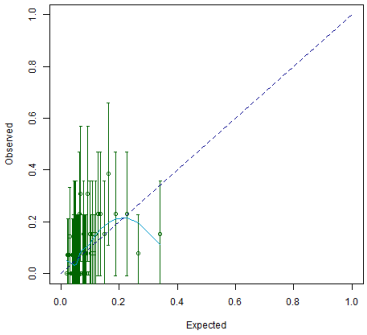


Figure B.9: IMD 4.

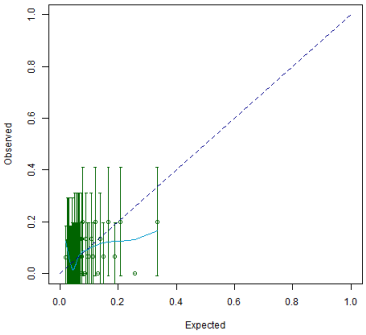


Figure B.10: IMD 5.

Table B.1: CTV3 Codes (n=45) used for labelling hip replacement (1).

CTV3 Code	Description
7K200	Primary cemented total hip replacement
7K20y	Total prosthetic replacement of hip joint using cement
7K20z	Total prosthetic replacement of hip joint using cement
7K210	Primary uncemented total hip replacement
7K21y	Total prosthetic replacement of hip joint not using cement
7K21z	Total prosthetic replacement of hip joint not using cement
7K220	Primary total replacement of hip joint
7K22y	Total replacement of hip
7K22z	Total replacement of hip
7K23.	Thompson hemiarthroplasty of hip joint using cement
7K230	Primary cemented hemiarthroplasty of hip
7K23y	Arthroplasty of hip joint using cement
7K23z	Arthroplasty of hip joint using cement
7K24.	Prosthetic uncemented hemiarthroplasty of hip
7K240	Primary uncemented hemiarthroplasty of hip
7K24y	Prosthetic uncemented hemiarthroplasty of hip
7K25.	Partial hip replacement by prosthesis
7K250	Partial hip replacement by prosthesis
7K25y	Partial hip replacement by prosthesis
7K25z	Partial hip replacement by prosthesis
X606J	Total replacement of hip
X606K	Partial hip replacement by prosthesis
XE08j	Total prosthetic replacement of hip joint using cement
XE08k	Primary cemented total hip replacement
XE08m	Total prosthetic replacement of hip joint not using cement
XE08o	Total replacement of hip
XE08r	Thompson hemiarthroplasty of hip joint using cement
XE08u	Partial hip replacement by prosthesis
XE2n7	Total replacement of hip
XS2Dh	Prosthetic uncemented hemiarthroplasty of hip
XaBFE	Prosthetic arthroplasty of the hip
XaBFG	Arthroplasty of hip joint using cement
XaBFH	Arthroplasty of hip without cement
XaBrw	Thompson hemiarthroplasty of hip joint using cement
XaF7j	Primary hybrid total replacement of hip joint
XaF7k	Primary hybrid total replacement of hip joint
XaF7l	Prosthetic hybrid total replacement of hip joint
XaMBd	Prosthetic hybrid total replacement of hip joint using cemented acetabular component
XaMBe	Primary hybrid total prosthetic replacement of hip joint using cemented acetabular component
XaMBj	Prosthetic hybrid total replacement of hip joint using cemented acetabular component

Table B.2: CTV3 Codes (n=45) used for labelling hip replacement (2).

CTV3 Code	**Description**
XaMBo	Primary hybrid total prosthetic replacement of hip joint using cemented femoral component
XaMBu	Prosthetic hybrid total replacement of hip joint using cemented femoral component
XaMC4	Prosthetic hybrid total replacement of hip joint using cement
XaMC5	Prosthetic hybrid total replacement of hip joint using cement
XaMCB	Prosthetic hybrid total replacement of hip joint using cement

Appendix C

Hip and Knee Replacement Prediction Appendix

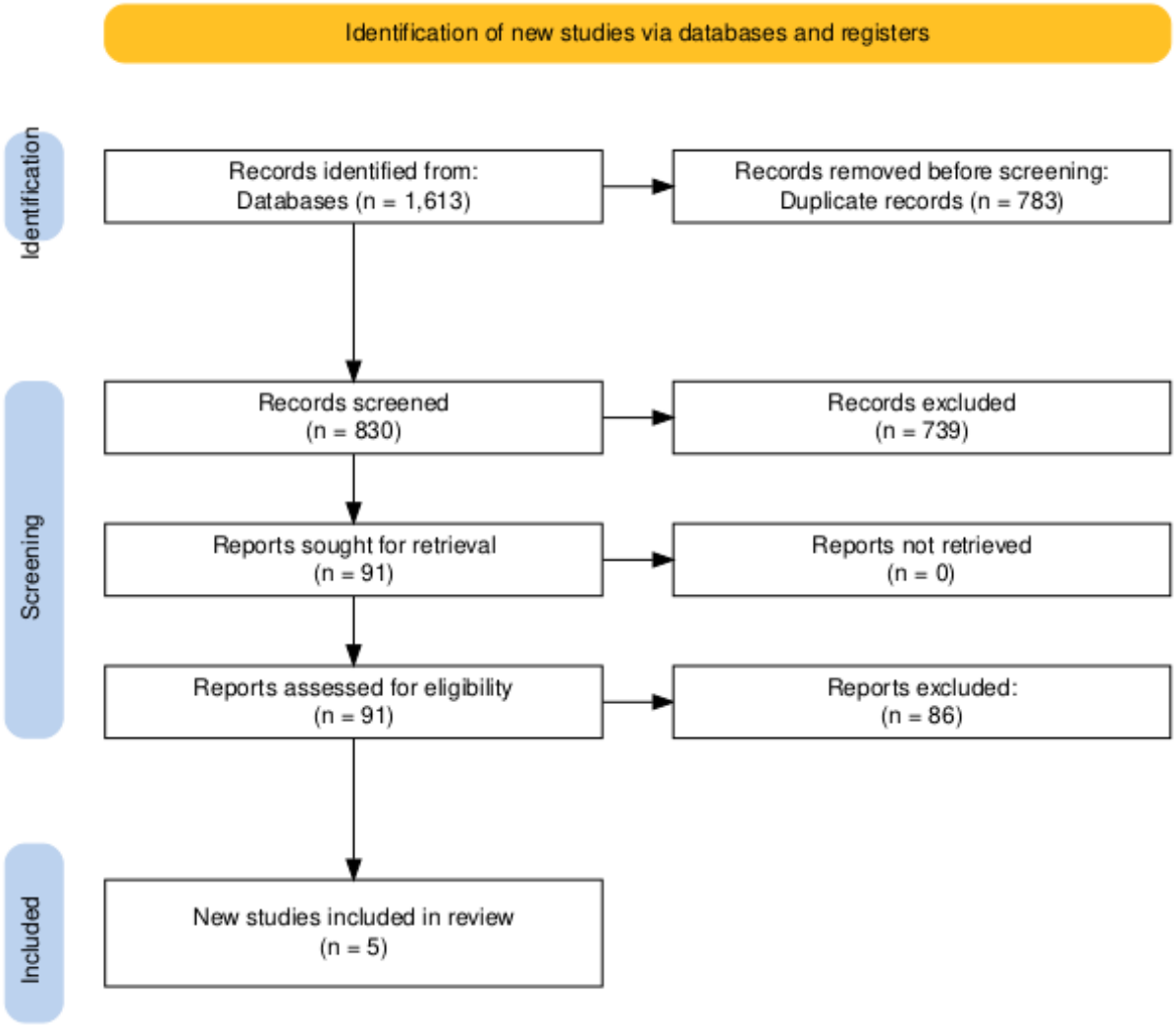


Figure C.1: PRISMA flowchart for systematic search of papers predicting hip and knee replacement risk using primary care data.

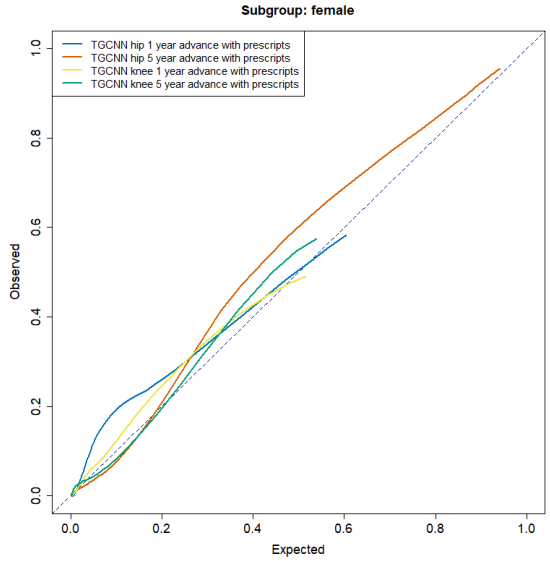


Figure C.2: Calibration curves for Females in each TG-CNN model.

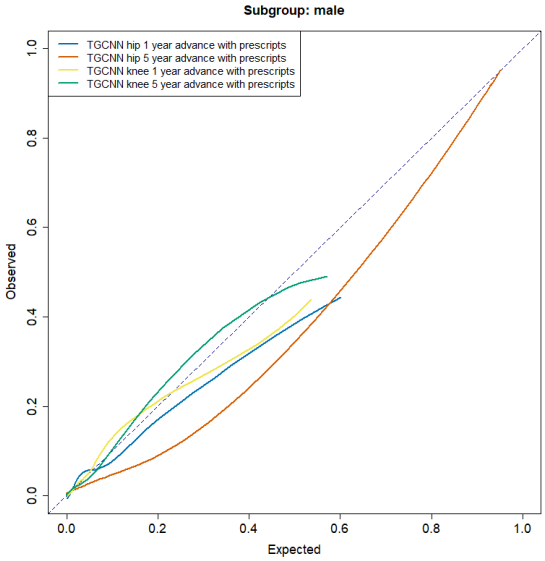


Figure C.3: Calibration curves for Males in each TG-CNN model.

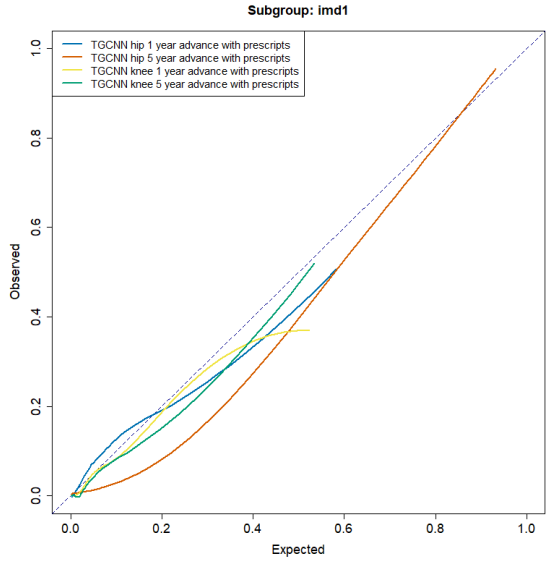


Figure C.4: Calibration curves for patients in the IMD 1 (most deprived) group in each TG-CNN model.

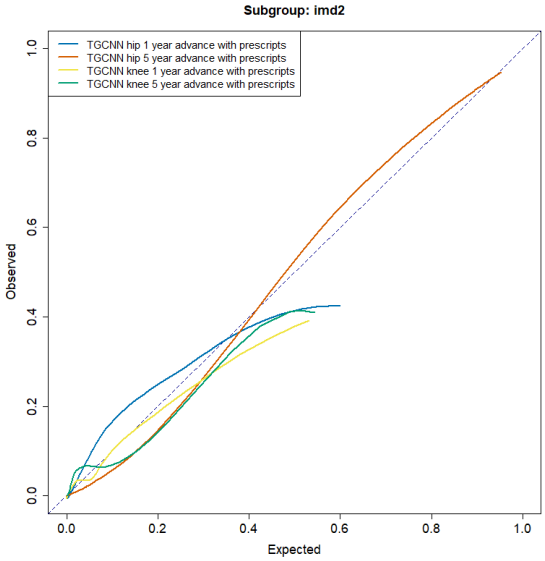


Figure C.5: Calibration curves for patients in the IMD 2 group in each TG-CNN model.

Table C.1: CTV3 Codes (n=33) used for labelling knee replacement.

CTV3 Code	**Description**
7K30.	Cemented knee arthroplasty (& total (& named variants))
7K300	Primary cemented total knee replacement
7K30y	Total prosthetic replacement of knee joint using cement OS
7K30z	Total prosthetic replacement of knee joint using cement NOS
7K31.	Arthroplasty knee joint without cement (& total)
7K310	Primary uncemented total knee replacement
7K31y	Total prosthetic replacement knee joint not using cement OS
7K31z	Total prosthetic replacement knee joint not using cement NOS
7K32.	Other arthroplasty knee joint (& total)
7K320	Primary total knee replacement NEC (& hybrid)
7K32y	Other total prosthetic replacement of knee joint OS
7K32z	Other total prosthetic replacement of knee joint NOS
7K37.	Cemented unicompartmental knee replacement
7K370	Primary cemented unicompartmental knee replacement
7K38.	Uncemented unicompartmental knee replacement
7K380	Primary uncemented unicompartmental knee replacement
7K39.	Hybrid unicompartmental knee replacement
7K390	Primary hybrid unicompartmental knee replacement
X606N	Arthroplasty of the knee
X606O	Prosthetic total arthroplasty of the knee
X606P	Prosthetic unicompartmental arthroplasty of the knee
X606Q	Prosthetic medial unicompartmental arthroplasty of the knee
XE07f	Knee arthroplasty (& replacement)
XE08w	Total prosthetic replacement of knee joint using cement
XE08y	Total prosthetic replacement of knee joint not using cement
XE090	Other total prosthetic replacement of knee joint
XE091	Primary hybrid total knee replacement NEC
XaBFJ	Prosthetic arthroplasty of knee
XaBFK	Arthroplasty of knee using cement
XaBFM	Arthroplasty of knee without cement
XaOPm	Unicompartmental knee replacement NOS
XaPtO	Hybrid prosthetic replacement of knee joint using cement
XaPtP	Primary hybrid prosthetic replacement knee joint using cement

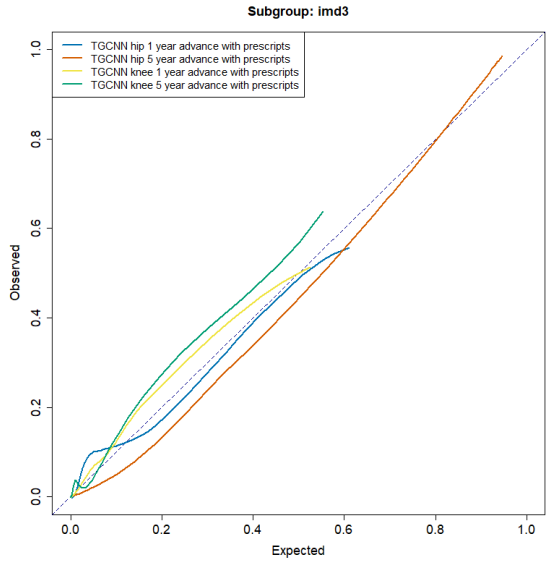


Figure C.6: Calibration curves for patients in the IMD 3 group in each TG-CNN model.

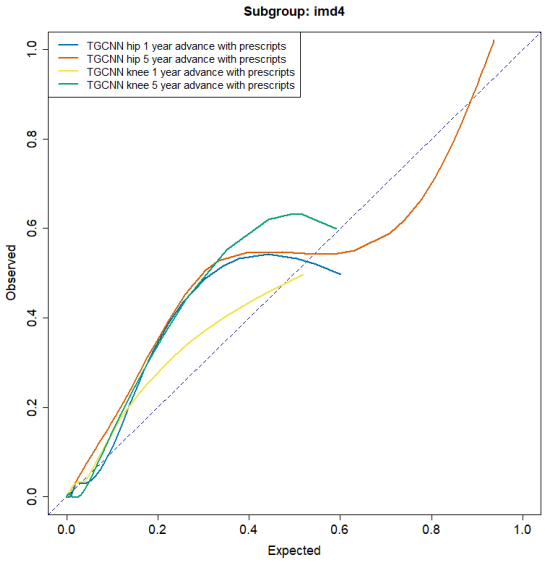


Figure C.7: Calibration curves for patients in the IMD 4 group in each TG-CNN model.

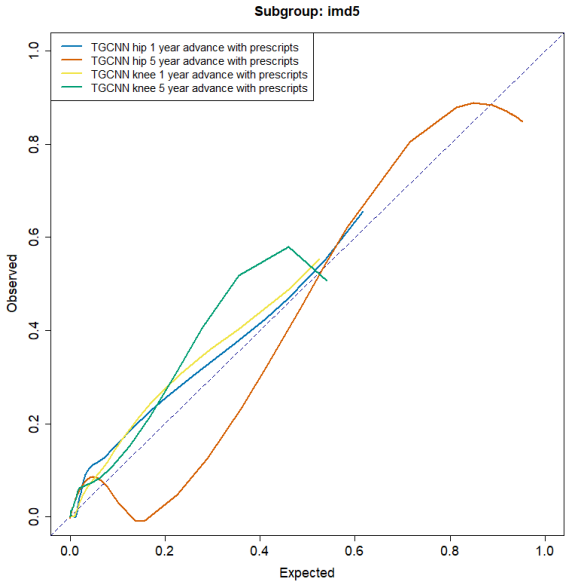


Figure C.8: Calibration curves for patients in the IMD 5 (least deprived) group in each TG-CNN model.

Appendix D

Explainability Appendix

D.1 Clinical Vignette

We gave clinicians the clinical example as given in Figures D.1 and D.2.

For each patient's graph set the following was asked: 1) Which of these visualisations is the easiest to interpret? 2) Do any of these visualisations not align with a trajectory you'd expect? 3) Does each graphing method effectively highlight the key factors influencing the model's prediction? 4) Any additional comments about the graph?

Then questions about the graphs overall were asked: 1) How useful are these graphing methods for aiding decision-making in a clinical-setting? 2) To what extent do the graphs support your understanding of the AI model's decision-making process? 3) Which of the four graphs do you prefer for visualising AI model predictions, and why? 4) Are there any additional features or information you would like to see included in these graphs?

Clinical Vignette – Example of How to Interpret Explainable Graphs

Scenario: a patient has presented themselves for a GP appointment and is having issues with hip pain and osteoarthritis. You think this patient might be at risk of needing a hip replacement in the coming years, but you would like to quantify that risk with more certainty using explainable artificial intelligence (AI) techniques.

Their clinical record is as follows:

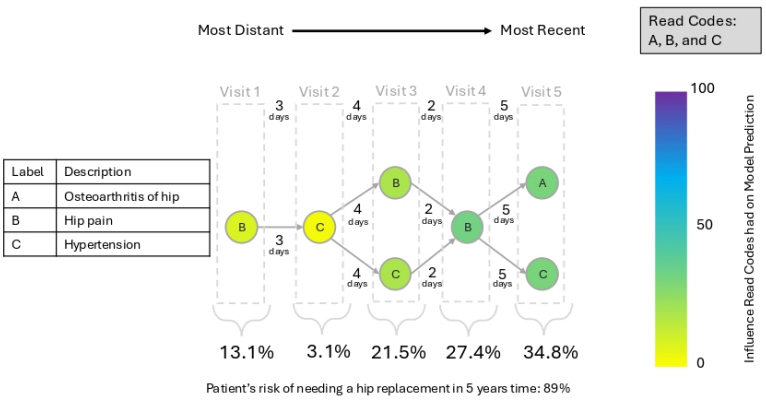
Date	Findings
01/08/2024	Hip pain
04/08/2024	Hypertension
08/08/2024	Hip pain Hypertension
10/08/2024	Hip pain
15/08/2024	Osteoarthritis of the hip Hypertension

We evaluate four different methods for visualising the features of Electronic Health Record (EHR) data that contribute to the hip replacement risk score generated by the AI model. Each method employs an interactive plot that enables users to navigate through a patient's history and identify the features (such as GP visits or Read Code pairs) that have the greatest impact on the model's predictions.

Method **a)** illustrates the influence of recordings from primary care visits on the model's prediction, focusing only on features that increase the risk score.

Methods **b)** and **c)** display the influence of these recordings on the model's prediction by considering both features that either increase or decrease the risk score.

Visualisation and labelling of methods **a**-**c**:



Explanation of methods **a-c** visualisation:

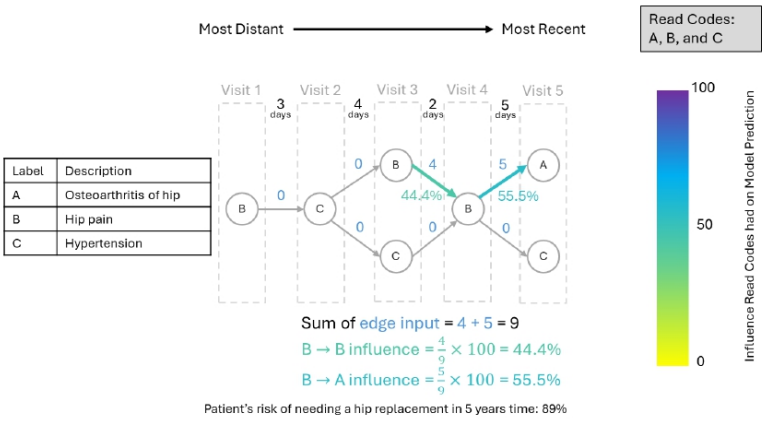
Visits 4 and 5 (the most recent GP visits) had the greatest influence on the risk prediction score (indicated by darker colours), with the recording of 'osteoarthritis of hip' having the most significant impact on the outcome. The patient has an 89% risk of needing a hip replacement

Figure D.1: Clinical vignette page 1.

within 5 years, and the model determined the extent to which each of these visits influenced the prediction.

Method **d)** illustrates the impact of pairs of Read Codes from the individual's EHR on the prediction outcome.

Visualisation of method **d**:



Explanation of method **d** visualisation:

The progression from a 'hip pain' recording in visit 4 to 'hip pain' in visit 5 accounted for 44.4% of the influence on the hip replacement risk. In contrast, the sequence of 'hip pain' in visit 4 followed by 'osteoarthritis of hip' in visit 5 contributed 55.5% to the hip replacement risk. The remaining interactions between Read Codes did not affect the risk score. With an 89% risk of needing a hip replacement within 5 years, the model determined that these two pairs of Read Codes influenced the risk prediction in the proportions stated.

Figure D.2: Clinical vignette page 2.