**Exploring safety, quality and resilience in health care**

**Allen Hutchinson**

**Thesis submitted for the degree of Doctor of Philosophy**

**by publication**

**School of Health and Related Research**

**The University of Sheffield**

**March 2014**

**Summary of Permissions to Publish**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Publications | Author-ship | Corresponding Author | Permission from corresponding author to publish | Permission to publish received from all co-authors | Publisher copyright policies |
| Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability.Quality and Safety in Health Care. 2006 | **1st** of 8 authors | Yes | N/A | Yes | Quality and Safety in Health Care -  Author can archive post print (final draft post-refereeing) |
| Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. Quality and Safety in Health Care. 2009 | **1st** of 7 authors | Yes | N/A | Yes | Quality and Safety in Health Care -  Author can archive post print (final draft post-refereeing) |
| Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway. BMC Health Services Research. 2007 | **Joint 1st** of 4 authors | No | Yes | Yes | BMC Health Services Research -  Author can archive publisher’s version/PDF |
| Note: Publication 4, referred to in the text, is not included in this eThesis because one author could not be contacted |  |  |  |  |  |
| Comparison of case note review methods for evaluating quality and safety in health care. Health Technology Assessment. 2010 | **1st** of 11 authors | Yes | N/A | Yes | Health Technology Assessment -  Author can archive publisher’s version/PDF |
| A structured judgement method to enhance mortality case note review: development and evaluation. BMJ Quality and Safety. 2013 | **1st** of 6 authors | Yes | N/A | Yes | BMJ Quality and Safety -  Author can archive post print (final draft post-refereeing) |

Publisher copyright policies are those available on the RoMEO publisher copyright and self-archiving database accessed 12/10/13 at <http://www.sherpa.ac.uk/romeo/search.php> A record of all co-author permissions is available to view in a separate file on request to the author at [allen.hutchinson@sheffield.ac.uk](mailto:allen.hutchinson@sheffield.ac.uk)

**Acknowledgements**

I wish here to thank those who have supported me through the process of writing this thesis.

My family have been supportive from the moment they heard that I was considering the submission of a thesis. When I have missed family events because of the task ahead it has all been taken in good heart. Very special thanks are due to my wife, Christine, for her constant support and encouragement.

Colleagues in ScHARR have also given me support and encouragement, especially those who have taken the PhD by Publication path before me. Finally, of course, I must thank Professor Suzanne Mason, who has been my mentor, supporter, reader of many drafts and generous of her time.

**Exploring safety, quality and resilience in health care**

**Abstract**

There still appears to be much to do to make the National Health Service in England a safer place for patients. Hospitals, in particular, are complex organisations in which staff and processes are under the twin simultaneous demands of an increasingly aged society and severe financial constraints. While much health care is well delivered, there remains a need to predict, and to explore, where and why problems occur. This thesis presents work which has refined methods and tools that can be used at health system and organisation levels to explore some key safety and quality issues in health care.

The six publications presented and discussed here were published during a seven year period between 2006 and 2013. They explore three important issues relating to safer health care – safety culture and incident reporting, prospective hazard analysis, and the use of improved case note review methods to evaluate the safety and quality of care in hospitals.

Two principal approaches to data access are presented in the publications. At the system and organisation level, information from large data sets was used to investigate the relationships between markers of safety and quality. At the health care provision level, data has been gathered about the work of health care professionals using mixed-methods approaches.

The publications are discussed across two inter-related concepts – healthcare safety and healthcare resilience. While the study of safer healthcare has a long history the concept of healthcare resilience is still being developed. Resilience is concerned with the way in which organisations and people can adjust and maintain their functioning in the face of challenge or adversity. Although the presented publications themselves do not explicitly consider research into resilience, this theme is used to reflect on the study results and their potential value to health services.

**Contents**

Abstract

1. Introduction………………………………………………………………….............8

1.1 The six presented publications…………………………………..............9

2. Considerations of healthcare safety and resilience.........................................10

2.1 The emergence of safer health care concerns……………….............10

2.2 Whose responsibility? From human error to organisational and human resilience…...............................................................................................11

2.3 A changing, and more positive, concept of safety…………................13

2.4 Presentation of the studies and publications.......................................13

3. Summaries of the presented publications, author contribution and impact of the studies..................................................................................................................15

3.1 Safety Climate and Event Reporting………..................…………........15

3.2 Prospective Hazard Analysis……........................................................18

3.3 Studies based on retrospective case note review................................20

4. Safety, quality and resilience in health care......................................................26

4.1 Safety Climate and Event Reporting....................................................26

4.2 Prospective Hazard Analysis................................................................30

4.3 Studies based on retrospective case note review................................33

5. Discussion.........................................................................................................42

6. Reflections.........................................................................................................44

7. Implications for practice and research...............................................................46

8. Conclusions.......................................................................................................48

9. References........................................................................................................49

10. Appendices

10.1 Publication 1. Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability......................................................57

10.2 Publication 2. Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System.................................................................................82

10.3 Publication 3. Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway............................................................................98

10.4 Publication 4. Assessing quality of care from hospital case notes: comparison of reliability of two methods...................................................117

10.5 Publication 5. Comparison of case note review methods for evaluating quality and safety in health care...............................................................118

10.6 Publication 6. A structured judgement method to enhance mortality case note review: development and evaluation.................................................336

**1. Introduction**

**1.1 Background**

This thesis introduces some of my work that was particularly focused on methods of evaluating the safety of care in UK health services and undertaken over a period of ten years. Much of the work relates to hospital care, although some elements include primary care contributions.

When there appear to be failures in care it has become the norm in the UK NHS to undertake ‘root cause analysis’ and often to seek out and improve points where staff may be seen to have made a personal contribution to failure. The research and analysis presented here takes a contrasting view. Conceptual models of health care accidents have changed over the past decade and failures in systems are now seen as the more likely causes of accidents.

While all of the research presented here relates to safety in systems, in this thesis I present the research and publications, where appropriate, in a more recent conceptual framework, that of organisational and human resilience (Hollnagel et al, 2006). Although healthcare resilience is still a developing science, I argue that the resilience approach is a more positive and just means to evaluate the safety and quality of healthcare. Further, the developing ideas on resilience fit well with recent thinking on the concept of safety that uses a more positive terminology in its definition.

In Section 2, I briefly discuss the rise in concerns over the safety of health care in the past two decades and consider some of the models which have been used to explore safety. This provides some of the policy and conceptual background for my work.

In Section 3, presents an overview of the six included published works, together with the methods used, my role in the studies and the impact the publications may have had.

Section 4 is an appraisal of the publications and the related studies in the context of literature on health care safety and resilience and Section 5 overviews the results of the analysis.

Section 6 is a reflection on my work, while Section 7 considers the implications that have arisen from the results presented in this thesis, from both a research and NHS practice perspective. Finally, Section 8 presents a short conclusion.

**1.2 The six presented publications**

All of the included publications have been peer-reviewed and are in the public domain through health services research journals. They are presented here in the order in which they are discussed.

**1) Hutchinson A**, Cooper KL, Dean JE, McIntosh A, Patterson M, Stride CB, Laurence BE & Smith CM. (2006)Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability.*Quality & Safety in Health Care*, 15 (5), 347-353.

**2)** **Hutchinson A**, Young TA, Cooper KL, McIntosh A, Karnon JD, Scobie S & Thomson RG. (2009) Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. *Quality and Safety in Health Care,*18, 5-10.

**3)** Dean JE, **Hutchinson A**, Hamilton Escoto K & Lawson R. (2007) Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway. *BMC Health Services Research,* 7, 89.

**4) Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Rantell K, Campbell MJ, Nicholl J & Irwin P. (2010) Assessing quality of care from hospital case notes: comparison of reliability of two methods. *Quality and Safety in Health Care*, 19:e2. doi: 10.1136/qshc-2007-023911.

**5) Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Young TA, Rantell K, Campbell MJ & Ratcliffe J. (2010) Comparison of case note review methods for evaluating quality and safety in health care. *Health Technology Assessment*, 14(10), 1-170.

**6) Hutchinson A**, Coster JE, Cooper K L, Pearson M, McIntosh A & Bath P A. (2013) A structured judgement method to enhance mortality case note review: development and evaluation. *BMJ Quality and Safety* 2013; doi:10.1136/bmjqs-2013-001839

**2. Considerations of healthcare safety and resilience**

**2.1 The emergence of safer healthcare concerns**

Although the concept of assessing care quality variation was first considered by Codman (1920) in the early 20th century, the need to focus on safer healthcare came to prominence more recently with the publication of studies on safety epidemiology from the United States (Brennan et al 1991) and from Australia (Wilson et al 1995). Study of failures in healthcare in Western health services suggests that about one in ten people who enter hospital care encounter some form of adverse event, with around one in 300 people suffering serious consequences ( Baker et al 2004, Zegers et al 2009, Hogan et al 2012). Research by Landrigan et al (2010) suggests an even higher rate of 25 harms per 100 admissions, a rate that appears to have failed to change over time in their study population.

This information led to the publication in the USA of ’To Err is Human’ (Institute of Medicine 2000), a seminal work which considered the impact of health care failures at the population level. It was followed in the UK by a policy document from the Department of Health – ‘An organisation with a memory’ (Department of Health 2000a) - which sought to raise safety awareness in the NHS and to plan a response to the emerging scale of the safety challenge.

At the same time, British public attention was also being drawn to variation in health care safety and the need to manage adverse events through the publicity surrounding the Bristol Royal Infirmary Inquiry (Department of Health 2001). The situation at Bristol Royal Infirmary was later described by Weick and Sutcliffe (2003) as a ‘culture of entrapment’.

Since the early research on adverse event rates there have been a number of other important inquiries concerning healthcare failures, concerning both the work of individual clinicians (for example, the inquiry into quality and practice in the NHS arising from the actions of Rodney Ledward) (Department of Health 2000b) and also of seemingly whole systems failures of care (for instance, the recent independent inquiry into care provided by Mid Staffordshire NHS Foundation Trust) (The Stationary Office 2103). These inquiries appear to show that the cultural and safety challenges for the NHS remain considerable at the time of writing.

Although hospitals are now beginning to be recognised as fast moving, complex and potentially unsafe entities, an initial challenge to organisational improvement was that the response to failures was perceived as a need to fix ‘human error’ in organisations that were erroneously considered to be safe environments (Woods 2006a). Thus accident models became centred on the role of humans, while root cause analysis and ‘blame and train’ (Cook et al 2008a) became established practice following an incident.

**2.2 Whose responsibility? From human error to organisational and human resilience**

Even as health services were rather belatedly awakening to the concept of ‘human error’, investigations into major industrial events such as the Piper Alpha explosion (Flinn 2006) were moving to identifying the place and role of individuals within the context of the working environment and business systems, rather than seeing the individual as the primary source of an event. Thus Perrow (1999) introduced the concept of the ‘Normal Accident’ (accidents are the product of normal operations) and Reason (1997) described the notion of organisational accidents. The conceptual perspective was moving from the individual as a cause of accidents to a more balanced view of the role of individuals as often being a key aspect in the prevention of failures and adverse events, or at least in ameliorating their effects rather than being the cause.

It no longer became useful, therefore, to ‘blame and train’, as seeing the human as part of the risk, but rather it was necessary to understand the underlying causes of unexpected (sometimes referred to as surprising) events. Woods and Cook (2002) laid out a framework for this new approach in their ‘Nine steps to move forward from error’ in which they critiqued the failure of organisations to take account of research evidence on systems failure and made the case for viewing people as making an important contribution to safety – as part of the solution. Weick’s (1987) thoughtful work on human reliability and the search for High Reliability Organisations (HRO) (Rochlin 1987) helped to transfer attention from the negative view on risks and risky organisations towards celebrating the ability of organisations and individuals to respond to unexpected challenges and to recover (Cook and Connor 2005).

Some have argued that the answer to the challenge of failure in complex healthcare situations is training for high reliability (Gaba 2003) and proposed using the HRO model found on some aircraft carriers or in the nuclear power industry. Reason (2000) saw advantages in the HRO model in these settings because such systems rehearse failure scenarios and put considerable resources into this effort. However Roberts et al (2005) pointed to the fragility of a High Reliability Organisation in healthcare, describing the development and subsequent demise of the HRO when key opinion leaders moved on. Furthermore, others have countered that healthcare is dissimilar from tightly managed industrial or military HRO settings and healthcare is instead confounded by complexity - ‘the enemy of safety’ (Woods et al 2010), and that sharp-end health care is confusing, ‘messy’ (Nemeth et al 2004, Cook et al 2008b) and ‘regularly irregular’ (Nemeth et al 2007). This picture of ‘messy-ness’ is well documented by Wears et al (2006) who have provided examples of Emergency Departments becoming overwhelmed by demand in. Similar situations are likely to be found in many other acute units in hospitals.

Resilience has been defined by Nemeth et al (2008a) as

‘the intrinsic ability of a system to adjust its functioning prior to, during or following changes and disturbances so that it can sustain required operations, even after a major mishap or in the presence of continuous stress’.

The concept of resilient healthcare organisations grew out of the recognition that it is normally not possible to build in the required redundancies in systems and staff that HROs require. But it should be the case that natural resilience could be enhanced in organisations to attempt to proof them, their teams and patients against unexpected events.

Hollnagel (2013) also suggests that performance variability – the factor that many healthcare systems try to control – is a normal feature of all complex systems and that eliminating variability is likely to reduce the effectiveness of the system.

In this context it has been argued that rather than being a vulnerability, human adaptability is a strength, contributing through human resilience to the resilience of the organisation as a whole (Reason 2008, Cook 2013). One of the useful adjunct approaches in this work is the study of the presence and effect of ‘gaps’ in healthcare (Cook et al 2000, Nemeth et al 2008, Patterson et al 2014) that can be used to identify vulnerabilities that might be fixed, or at least bridged, when seeking to enhance the resilience of a team or system.

Increasingly, detailed observational research is documenting resilient responses at the organisational level (for example, Cook and Connor 2005). Nemeth et al (2008) and Wears et al (2006) have used workflow analysis during patient admission and handover to explore the settings in which clinical decisions may have to be made under stress and thus to study the resilience of the workforce in overcoming institutional barriers to better care.

At the individual level, healthcare professionals are shown on a regular basis to be fixing systems that do not work properly (Smith et al 2013). Or in other words, they can act in a positive, resilient manner – ‘taking it in one’s stride’, as Cook and Nemeth (2006) so aptly describe it. Interestingly, some of the examples of ‘fixes’ so far published in the healthcare literature, for instance by Smith et al (2013), appear to be similar to what one might call everyday clinical practice.

Recent work has also explored the concept of resilience markers which might act as an early warning or as a pointer towards examples of good practice or where gaps in care may indicate potential difficulties or failures. Drawing on experience in the nuclear industry, Furniss and colleagues (2011a) have proposed a resilience markers framework to identify and categorise resilient actions in healthcare (Furniss et al 2011b), although this approach is still in the early stages of development. Indeed, Cook (2013) points out that the work is challenging and that our knowledge of resilience in all fields is based on a few examples rather than having a well worked through conceptual base. Thus the notion of resilience in healthcare is still ‘work in progress’ but provides a valuable perspective through which to explore safer care initiatives.

**2.3 A changing, and more positive, concept of safety**

It seems to have been a challenge for healthcare in general to come to a useful – that is, understandable and workable – definition of safety. By conceptualising safety as a ‘dynamic non-event’ Weick (1987) helped to show that the idea of safety might be more complex than a definition based on the notion of freedom from events. But how might this new definition be accomplished, particularly when others were seeing accidents in large organisations as ‘normal’?(Perrow 1999)

My own work on reviewing the care of people who died in hospital has shown that many people who have apparently poor outcomes can be recipients of care that is both of good quality, sometimes under difficult conditions, and is safe (even event free) (Hutchinson et al 2013). And in a UK healthcare service where ‘efficiency savings’ mean care must be provided in a ‘faster, better, cheaper’ mode, I have until recently found it difficult to identify a definition of safety that encompasses the complex reality of health care, the need to reduce adverse events and the obvious commitment of most healthcare professionals.

Thus the emergence of new ideas on safety, in which (so-called) Safety II is defined as -

‘the ability to succeed under varying conditions, so that the number of intended and acceptable outcomes is as high as possible’ (Hollnagel 2013),

- for me allows a new paradigm for both research and practice, and one in which to review the potential impact of the research presented here.

**2.4 Presentation of the studies and publications**

I have therefore set out my work presented in the six publications in the context of the recent thinking on safer healthcare and on resilience. The studies are primarily methodological in nature but with UK NHS applications. For example, at the time of writing, *Publication 6* is informing a large NHS safety and quality initiative. Because there are either conceptual or methodological linkages between the studies on which the publications are based I have chosen to group the publications together, and to discuss them, under three headings:-

(**1**) Safety Climate, Safety Culture and Event Reporting,

(**2**) Care Pathway Prospective Hazard Analysis,

(**3**) Studies based on Retrospective Case Note Review.

**3. Summaries of the six presented publications, author contribution and impact of the studies.**

This section comprises brief summaries of the six published research papers on which this thesis is based and they are grouped together under the three headings as outlined in Section 2. Full copies of the papers are in Appendices 10.2 to 10.7.

My contributions to Publications 2, 3, 4, 5 and 6 are included in the format which was required by the publishers, agreed with co-authors and printed in the journals. There was no journal requirement for a contribution statement for Publication 1 and I have provided an explanatory statement for this. For Publication 3 I was not the corresponding author but the print version of the publication acknowledged that I had made an equal level of contribution to the corresponding author. Here I have included the published contribution statement together with an additional explanatory note of my role.

The impact of each publication is also considered here. Taken together the published papers represent a multi-methods range of work round the central theme of assisting healthcare professionals to evaluate the safety of the service within which they work. The research explores aspects of safer health care in hospitals and at the boundary between primary and secondary care. Each of the studies has provided information that has been useful to both the research community and to health providers. An impact statement accompanies each of the presented publication summaries.

**The presented publications**

***3.1 Safety Climate and Event Reporting***

Two publications are presented here, linked conceptually to safety culture, since one study is directly about safety climate, while the second study found that higher trends in reporting adverse events were linked to a more positive safety culture and there is a relevant literature on this subject.

|  |
| --- |
| *Publication 1*.  **Hutchinson A**, Cooper KL, Dean JE, McIntosh A, Patterson M, Stride CB, Laurence BE & Smith CM.(2006)Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability.*Quality & Safety in Health Care,* 15, 347-353 |

The study reported in *Publication 1* explored the appropriateness of using one of two selected, US designed, short safety climate questionnaires in the UK National Health Service (NHS). A small group of front line staff completed both questionnaires and then discussed their understanding of, and their responses to, the questions with a research team member. These results were used to select a safety climate questionnaire which best fitted the respondents’ understanding of the questions - the 27 item Teamwork and Safety Climate Survey (Sexton et al 2003, The University of Texas Center of Excellence for Patient Safety Research and Practice 2005).

In the second phase, postal questionnaire responses to the survey from 897 clinical staff, working either in primary care or secondary care, were analysed using exploratory and confirmatory factor analysis.

Results showed that there were differences between the factor structure based on US hospital healthcare workers and the factor structure based on data from UK healthcare staff. Additionally, there were some differences in teamwork climate between the UK staff groups (better in primary care than in secondary care). The mean scores for the three safety climate factors were also lower for the secondary care subsample.

***Contribution statement***

There was no contribution statement published with this paper.

I was the originator of the idea for this study, which arose from my study of safety climate instruments that were currently in use, or being promoted, in the US health care sector in 2002. I wrote the research application, obtained the project funding, gained research ethics approval and led on discussions with four local NHS hospitals and 13 Primary Care Trusts.

I set out the research plan, identified the required skills and brought the research team together, directing the project on a day-to-day basis. I led all of the research team meetings and made contributions to data acquisition, management and analysis.

I wrote the final report to the funding body. In collaboration with other team members, I was the lead author of the drafts and managed the review and publication process.

***Impact***

This study noted that ‘there are enough cautionary points arising from the item content and the factor analysis of this questionnaire to suggest there is more to be done in exploring the properties of safety climate instruments’, especially those developed in another health culture. Similar concerns have been raised in subsequent studies (Waterson et al 2009). The assessment of safety climate continues to be important internationally and this study has been cited in 42 publications.

|  |
| --- |
| *Publication 2*.  **Hutchinson A**, Young TA, Cooper KL, McIntosh A, Karnon JD, Scobie S & Thomson RG. (2009) Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. *Quality and Safety in Health Care,* 18, 5-10 |

*Publication 2* reports a study in which the patterns of reporting healthcare incidents by 148 acute hospitals in England to the National Patient Safety Agency (NPSA) were examined to determine whether there were any defining characteristics of hospitals which had lower or higher reporting rates. The only resulting trend of note was that hospitals with higher reporting rates had a lower overall percentage of ‘slips, trips and falls’.

Responses from the NHS Annual Staff Survey (Healthcare Commission 2006) were used to explore any relationship with the reporting trends of the hospitals. There were a number of significant linear relationships between higher reporting-rate hospitals and NHS Staff Survey responses on fairness and effectiveness of reporting, encouragement to report and on the question on blaming and punishing for making errors. Additionally there was a significant positive correlation between higher reporting rates and a higher proportion of staff having reported the last error/near miss they saw. There were no significant correlations observed with Hospital Standardised Mortality Ratios (see also the results of *Publication 5*) or with a group of Patient Safety Indicators (Agency for Healthcare Research and Quality 2006).

***Published contribution statement***

AH wrote the research proposal, directed the study and contributed to analysis and writing, TY undertook much of the analysis and contributed to the writing, KC managed the research process, undertook some analysis and contributed to the writing, AMc undertook the qualitative evaluation and contributed to the findings, JK contributed to the design and analysis, SS and RT proposed the project and contributed to design, writing and interpretation.

***Impact***

Incident reporting in healthcare remains a challenge in many countries and a significant international literature has built since the publication of this paper. Human factors, including safety culture issues, continue to be seen as having an impact on incident reporting and may be important as one of the elements of a resilient organisation. This publication remains of interest to other researchers and has been cited in 44 other publications.

***3.2 Care Pathway Prospective Hazard Analysis***

|  |
| --- |
| *Publication 3*.  Dean JE, **Hutchinson A**, Hamilton Escoto K & Lawson R. (2007) Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway. *BMC Health Services Research,* 89  <http://www.biomedcentral.com/1472-6963/7/89> |

This study explored the possible use of failure mode and effects analysis in the NHS, as a means of assessing risks in newly established care pathways. A combination of evaluation methods were applied to a recently established chronic obstructive pulmonary disease (COPD) care pathway which bridged between the hospital, community care and the patient’s home.

Using interviews with staff and patients and observation of care pathway processes, both in hospital and in the community, care pathway flowcharts were drawn, critiqued and reconfirmed with staff. The interviews and the flowcharts identified possible hazard areas and a list of these was created, and ranked by staff in order of severity, using a Delphi process (Hasson et al 2000). In the final stage, a small group of staff, led by me as facilitator, used healthcare failure mode effects analysis (HFMEATM) (De Rosier et al 2002) to analyse two problem areas – communication with primary care teams and communication with the hospital bed bureau.

The study illuminated the complexity of what was thought to be a relatively straightforward implementation of a new service for patients, in which little consideration had been given to the effectiveness, safety and resilience of the system once in place – and no previous attempt had been made to evaluate problems. Findings also revealed, however, that the effort, complexity and time involved in undertaking failure mode and effects analysis was considerable.

***Published contribution statement***

AH conceived of the study, obtained funding, oversaw the running of the study, analysed data and contributed to the development of the care pathway and the writing of the manuscript. JED obtained ethics approval, developed data collection materials, conducted fieldwork, analysed data and contributed to the development of the care pathway and the writing of the manuscript. KHE contributed to the literature searches, development of the care pathway, development of interview schedules and the writing of the manuscript. RL was involved in the conception of the project, project meetings, data collection and contributed to the development of the care pathway and the writing of the manuscript.

Additional contribution commentary

Although not the corresponding author on *Publication 3*, my contribution to the research and writing was equivalent to that of J.E. Dean and we are identified in the publication as such. Both JED and I each contributed about 45% of the total input to the project.

Following discussions with the Health Care Failure Modes and Effect Analysis (HCFMEATM) design team I decided that this was worth testing in UK healthcare. I designed a multi-step process for identifying the complexities of a recently established ‘revolving door’ system for ensuring fast hospital access for people with chronic obstructive pulmonary disease. I used my experience of care process mapping, the Delphi method, direct observation and interviews with staff and patients, to provide the evidence. I developed the detailed methods and the frameworks for each step of the analysis. JED undertook the fieldwork, Delphi process and data management. I led the analysis of each step of the project, working jointly with JED and chaired the HCFMEATM session. I wrote the final report for the funding body and led on all of the drafts of Publication 3 up to the point where publication had been rejected by two journals. JED led the third, successful, drafting and journal submission process.

This additional statement has been agreed with JED, the corresponding author.

***Impact***

Prospective hazard analysis in UK healthcare is still only undertaken on a limited scale. The study findings showed that the mixed-methods approach can identify potential failure points in a care pathway that bridges across hospital and community care. But so far as NHS practice is concerned, the requirements of expertise and staff time commitments are such that prospective hazard analysis might be best restricted to high risk/high intensity systems. More recent work from Ward et al (2010) supports these findings. Citations to date,10.

***3.3.3 Studies based on retrospective case note review***

*Publications 4, 5 and 6* all arose from one major investigation based on retrospective case note review in hospitals, conceived and directed by the candidate.

*Publication 4* reported first on the comparative effectiveness of two methods of retrospective case note review:- explicit criterion based review and structured implicit review. Some of these results were subsequently included in Chapter 2 of *Publication 5* for completeness, as was required by the research contract.

*Publication 5* is extensive and reports on two separate studies – (i) the case note review comparison and (ii) an exploration of the relationship between safety and quality markers and proxy measures of outcomes in twenty hospitals.

It is used here for two purposes. First, because it contains the study (reported principally in Chapter 3 of the publication) which explores the relationship between the safety and quality of care assessed from case notes and hospital-level outcomes of care. It also contains, in Chapter 2, additional evidence on the relationship between hospital mortality levels for the study indicator conditions and safety and quality of care assessed from the case notes. Second, *Publication 5* contains much of the background methodological information that *Publications 4* and *6* only report in brief.

*Publication 6* reports on an analysis of the qualitative, judgement-based data collected as part of the process - outcomes component of the research. It develops the assessment process into the structured judgement review method as applied to people who die in hospital.

|  |
| --- |
| *Publication 4*.  **Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Rantell K, Campbell MJ, Nicholl J & Irwin P.(2010) Assessing quality of care from hospital case notes: comparison of reliability of two methods. *Quality and Safety in Health Care*, 19, e2. doi: 10.1136/qshc-2007-023911. |

This stage of the study was undertaken to explore and refine the effectiveness of retrospective case note review methods. *Publication 4* sets out the key results from the comparison of reliability between two recognised methods of case note review – explicit criterion based review and implicit structured review.

The criterion based method was similar to that used in national clinical audit programmes and comprised a specific set of evidence-based data to be captured by the reviewers. The implicit review was a development of the initial structured implicit review method developed by RAND (Rubenstein 1991). To the RAND method was added a phase of care framework, quality scores and explicit written judgements on the safety and quality of care - both good and poor. Thirty-nine reviewers undertook 1473 implicit reviews and 1389 criterion based reviews in nine randomly selected acute hospitals. Two or three reviewers undertook a review of each set of clinical notes, using both criterion-based review and implicit review.

Analysis of the data was undertaken to explore which methods were most reliable in the hands of which type of staff, using intrarater consistency, interrater reliability between pairs of staff using intraclass correlation coefficients (ICCs), and between-staff comparison. For criterion based scores, inter-rater reliability was moderate to good. Within each staff group using implicit review there was moderate inter-rater reliability (ICC 0.46 to 0.52).

For criterion-based clinical audit of the type often used at a national level, results suggested that clinical audit staff could do the work as well as the more expensive nurses and doctors. Whereas for the more complex work of detailed assessment of safety and quality of the care the clinical staff showed a higher reliability and doctors provided more explicit information than did the nurses.

***Published contribution statement***

AH developed and directed the study and acted as lead author for this paper. JED project managed the study, collected and analysed the data. KLC assisted with project management, undertook recruitment and data analysis. AMc acted as senior methodologist and lead qualitative researcher. SJW lead the statistical analysis and contributed to methods development. PAB contributed to methods development and analysis. MP acted as senior clinician, contributing to recruitment, methods development, analysis and writing. KR undertook statistical analysis and contributed to methods development. MJC acted as senior statistical adviser. JPN contributed to the development of the study and to the analytic framework. PI contributed to methods development and to recruitment. All authors contributed to the writing of this paper.

***Impact***

Retrospective case note review remains the main approach to assessing safety and quality of care of people admitted to hospital. In order to understand variations in safety and quality however, structured implicit review proved to be both feasible and to have reasonable inter-rater reliability between physicians. This study has implications for any healthcare system where case note review is being undertaken and the methods have, for example, been used in the UK to assess the safety of emergency medicine practice (O’Hara et al 2010), Citations, 4.

|  |
| --- |
| *Publication 5*.  **Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Young TA, Rantell K, Campbell MJ & Ratcliffe J. (2010) Comparison of case note review methods for evaluating quality and safety in health care. *Health Technology Assessment* , 14(10), 1-170. |

The work on process-outcome comparisons is drawn primarily from Chapter 3 of the publication. For this study, data collection was undertaken in 20 acute hospitals in England during 2006-7. Forty physician reviewers used a combination of criterion based review and structured implicit review, providing written commentaries on the quality of care and associated quality of care scores, on a total of 1565 sets of case notes (873 from people with chronic obstructive pulmonary disease and 692 from people with heart failure). Scores and commentaries from the implicit reviews were provided for three phases of care and for care overall.

Only the quality and safety scores and the criterion based scores were used to provide the healthcare process data in this analysis. None of the judgement commentaries were analysed in this publication.

Outcome data was composed of hospital level variables from a number of sources such as Dr Foster Hospital Standardised Mortality Ratio (HSMR), Hospital Episode Statistics for chronic obstructive pulmonary disease and heart failure, Health Care Commission targets and data from the NHS Staff Survey. There were some commonalities with this data set and that used in *Publication 2* above.

There were only limited correlations between hospital mortality rankings and the overall implicit care scores for both chronic obstructive airways disease and heart failure and there were no correlations with HSMR data.

***Published contribution statement***

AH, principal investigator, principal author of application, overall management of project, main author of report. JEC, project manager, contribution to application, major contribution to fieldwork and analysis. KLC, project management, major contribution to fieldwork and analysis. AMc, lead qualitative researcher, major contribution to qualitative analysis and significant contribution to project development. SJW, lead statistical advisor and senior statistical analyst. PAB, senior contributions to project development and analysis. MP, provided expertise in quality assessment methods and lead clinical advisor. TAY, statistical analyst taking lead in the development and analysis of the outcomes component of the study. KR, statistical analyst making contributions to the management of data sets in the first stage of the study. MJC, provided senior and specialist statistical advice to the project design and analysis. JR, provided health economics advice and analysis.

***Impact***

The process-outcomes study failed to identify a clear link between the safety and quality of care identified by structured implicit case note review and other proposed markers of safety. Nevertheless, from a methodological viewpoint, the study did raise the question of whether mortality scores based on statistics have any value. It did break new ground at an international level in refining structured explicit judgement case note review. Citations, 4.

|  |
| --- |
| *Publication 6*.  **Hutchinson A**, Coster JE, Cooper K L, Pearson M, McIntosh A & Bath P A. (2013) A structured judgement method to enhance mortality case note review: development and evaluation. *BMJ Quality and Safety,* doi:10.1136/bmjqs-2013-001839. |

This publication presents an analysis of the structured judgement comments relating to the care of the 119 people who died within the cohort of 1566 for whom structured implicit review was undertaken in the process-outcome study (reported in *Publication 5*).

The analysis explored whether physician reviewers were able to consistently provide short, structured judgement comments on quality of care that they could also justify with an appropriate care score. Using an analysis framework for the textual data initially proposed in *Publication 5*, all of the structured judgement comments provided by the reviewers were analysed by me to determine whether they were implicit or explicit, and whether they were commenting positively or negatively on care.

Physician reviewers were on the whole able to make informative clinical structured judgement comments across all of the phases of care and usually provided a coherent quality of care score. The majority of comments (83%) were explicit judgements. As a result of this analysis it was proposed that the method could be used to evaluate deaths in hospital, in a way that would allow rapid review of the safety of care for people who died.

***Published contribution statement***

AH took the lead in the conception and design of the study, took the lead on the analysis of the qualitative mortality review data, was principal author of all of the drafts of this paper. JEC contributed to the conception and design of the study, undertook data collection and analysis of the mortality review data. KLC made contributions to the design of the study, undertook data collection and contributed to analysis of the mortality review data. MP contributed to the conception and design of the study overall and contributed to the interpretation of the mortality review data. AMcI contributed to the conception and design of the study overall and took a lead on the qualitative analysis framework. PB contributed to the conception and design of the study and qualitative analysis framework and undertook the statistical analysis for the quantitative analysis. All authors contributed to all of the drafts of this paper and have given approval for this version of the paper to be published.

***Impact***

This is the most recent publication presented (2013) and there are as yet no citation records available. From a clinical practice perspective, however, there is already evidence that the work is valued, since the explicit structured judgement case note review method has already been included in the NHS Yorkshire and the Humber Mortality Review programme.

**4. Safety, quality and resilience in health care**

In this section I review and appraise the six presented publications within the context of the new, more positive, concepts of healthcare safety and, where appropriate, from the perspective of healthcare resilience. Because the work explores a number of fields of healthcare safety, this retrospective draws on the literature on safety and resilience more broadly than just on the specific subject areas of the six publications. Rather than being a systematised review it seeks to place each of the presented publications in the context of the literature on the recent safety and resilience paradigms.

**4.1 Safety Climate, Safety Culture and Event Reporting (Publications 1&2)**

***Safety climate evaluation (Publication 1)***

Healthcare came rather late to the field of safety climate assessment and initially drew heavily on safety climate instruments from other industries because there was already a literature from which safety dimensions could be drawn be useful in healthcare assessment.

Flin and colleagues (2000), working in the petrochemical industry and subsequently in healthcare, described safety climate as the ‘surface features of the safety culture discerned from the workforce’s attitudes and perceptions at a given point in time’. When the study associated with *Publication 1* was undertaken, research by Helmreich and Merritt (1998) had derived a number of important safety climate factors from earlier research on aviation, and had developed measures for use in healthcare organisations. These included the 65 item Safety Attitudes Questionnaire (SAQ) and the shorter Team Work and Safety Climate questionnaire (TWSC) (Sexton et al 2003). In a similar vein, Singer et al (2003) developed the Patient Safety Cultures in Healthcare Organisations survey (PSCHO) from five other questionnaires, including one from the US Naval Aviation service. The PSCHO was then used by them to survey staff in 15 US hospitals.

The psychometric properties of the TWSC and the PSCHO questionnaires had not been published when I devised the study in 2002. Nevertheless, it was determined that these two questionnaires were the best available for our purpose and, of the two, the TWSC questionnaire was chosen as being more understandable to UK users, although it also required amendments to the wording of the questions.

We found that the TWSC questionnaire could be used among UK healthcare staff but that the item content and factor structure required revision. The key messages from this study were that is was necessary for UK users to take care when using questionnaires from another healthcare setting, such as US hospital care, both because the factor structures may not remain robust when applied to UK data and because the language can be subtly different. This remains the case in more recent times. Currently available safety climate questionnaires are much improved - for example the Hospital Survey of Patient Safety Culture (HSPSC) (Nieva and Sorra 2003) was specifically developed for healthcare. Nevertheless, Waterson and colleagues (2010) also found differences with factor structures using the HSPSC when comparing the results from US respondents and UK respondents, and our own work with the HSPSC found factor structure differences using data from Saudi Arabia (Alonazi et al 2011).

***Event reporting (Publication 2)***

Reason (1990) identified four key reporting-related components of an effective safety culture – that event reporting should

‘be valued, just, flexible and should encourage learning’.

In a discussion on aviation safety, Eiff (1999) noted his view that

‘one of the foundations of a true safety culture is that it is a reporting culture’

and that one component of a good reporting culture is

‘the free and uninhibited reporting of safety issues that come to the attention of employees during the course of their daily activities’.

While healthcare event reporting systems are now quite widespread, a number of authors have been critical of the implementation process, suggesting that the human factors in reporting and transmitting adverse event reports can be a challenge (Battles and Stevens 2009). Shojania (2008) reported his experience of reporting systems being influenced by personal choice – to report or not to report – and also on the variations of reporting rates within and between hospitals. Both internal and external factors might influence reporting rates, with a tendency of external agencies to regard low levels of reporting as being a positive marker when the opposite might be the case. A high reporting rate might indeed be a marker of a fit-for-purpose safety culture, although process issues such as failure to use agreed data definitions are still leading to concern over the value of reporting rates as effective benchmarks (Tanner et al 2013).

The response of employers to the reporting of incidents may be critical in determining whether a reporting rate is telling an accurate story about the institution. Both Shojonia (2008) and Firth-Cozens and colleagues (2004) identified the lack of response to reported incidents from senior hospital management as a major barrier to encouraging staff to report. Benn et al (2009) explored the issue of reporting barriers, finding that

‘risk reporting and feedback systems are highly variable (between hospital trusts) in terms of coverage of reporting and feedback’.

*Publication 2* investigated associations between the rate at which English hospitals were reporting to the recently established National Reporting and Learning System (NLRS) and other indicators that were conceptually linked with safer care. Although a wide range of indicators were accessed the study found only a few associations between hospital adverse event reporting rates and safety indicator data. There was, however, a clear positive association between some safety culture items and the rate of reporting. This association is in the same expected direction as that proposed by Eiff (1999) - better safety cultures promote event reporting. Reason (1997) regarded a reporting culture as one of the four key aspects of a safety culture, linking the willingness and ability to report with the presence of a just culture, which itself requires an input from senior leadership.

Research on the effectiveness of Leadership Walkarounds (Frankel et al 2003), and also by myself and colleagues (AlBuharan et al 2008), suggests that senior leadership is indeed important in a complex environment where safety is always a concern. Benn et al (2009) also see visible senior leadership as an important aspect of the development of safety culture.

The importance of engaged leadership in promoting safety is supported by the work of Edmondson (2004) and of Tucker (2003), who have explored the theme of why hospitals fail to learn from failure. Both authors suggest that an organisation’s ability to learn from failure (that is, its resilience) is measured by its ability to handle both large and small events.

Edmondson (2004) also argues that hospitals have not learned from failure because of a poor climate of reporting mistakes or problems and that-

‘features of work design and culture of most hospitals make workarounds and quick fixes the dominant response to failure’

-rather than there being any systematic response to the problems being reported.

Results in *Publication 2* indicate there were some positive associations between reporting rates and NHS 2004 Staff Survey questions, such as encouragement to report, and there were significant linear relationships between higher reporting rates and a higher proportion of positive responses.

These limited positive findings suggest that, had it been possible, it might have been valuable to have used the results of a full safety climate questionnaire in the study. However, there remains a tendency in the NHS to use individual safety climate questions as indicators and the NHS staff survey provides only a limited set of items. Perhaps this is reasonable. As Akselsson and colleagues (2009) subsequently suggested, the compromise solution of using shorter questionnaires is to use individual items ‘sometimes pooled into aspects making sense in practice’.

What is the place of resilience in the exploration of healthcare safety climate and culture? Flinn (2006) points out that whether or not managers/professionals make resilient sacrificial decisions in times of organisational stress (as exemplified by the work by Wears et al (2006) on Emergency Departments going into ‘free fall’) depends not only on staff skills and commitment but also on the general organisational level of commitment to safety. In a resilient organisation staff must be able to speak up when they meet problems and safety climate surveys are one means of measuring both managerial and clinical staff safety resilience.

Furthermore, when taking resilience into account, it also needs to be recognised that a poor organisational culture can have a negative influence on the safety of care. Weick and Sutcliffe (2003) reviewed the events surrounding the Inquiry into children’s deaths during cardiac surgery at the Bristol Royal Infirmary (Department of Health 2001). They describe the overall picture at the hospital as

‘a culture of entrapment’ - ‘the process by which people get locked into lines of action, subsequently justify those lines of action, and search for confirmation that they are doing what they should be doing’.

Additionally, in an exploration of the socio-cultural aspects of improving safety for patients, Waring (2013) points to the

‘complex links between safety, identity and professionalism’.

Perhaps in recognition of these negative elements, Akselsson et al (2009) propose, from an aviation industry perspective, that resilience assessment should be a considered part of safety culture assessment. The aim would be to mitigate pockets, perhaps whole systems, of poor and unsafe culture that might be found, by using workforce sub-group population analysis of safety climate questionnaire results. Our study did show safety climate differences between professional groups, although the number of respondents was quite high and smaller sub-group differences in a single institution may be more difficult to detect using a questionnaire approach.

Nevertheless the challenges of interpretation should not stop the use of safety culture assessment as part of the process of exploring resilience. Safety climate could be viewed at the institutional, group or even ward or operating theatre level as a marker or indicator of resilience, much in the way that Furniss and colleagues (2011a, 2011b) have proposed in their layers of resilience markers. Further research is needed to explore the value of individual safety climate questions acting as resilience markers.

**4.2 Prospective hazard analysis (Publication 3)**

Searching prospectively for hazards in health care offers the opportunity of revealing latent safety vulnerabilities and presents the possibility of making changes before failures occur. In principle, therefore, there are advantages in applying such frameworks in health care, both for complex technical situations such as surgical sterilisation procedures (Linken et al 2005) and in longitudinal processes such as medication administration (Karnon et al 2007).

Failure mode and effects analysis was previously proposed as a means of identifying hazards in medication administration processes (Cohen et al 1994, McNally et al 1997). Subsequently the US Veterans Health Administration refined this method (Healthcare Failure Modes and Effects AnalysisTM (HFMEATM)) (DeRosier et al 2002), including care process mapping to identify failure points.

Wreathall and Nemeth (2004) explored the potential role of Probabilistic Risk Assessment (PRA) as a complement to FMEA and root cause analysis because it can

‘take account of the more complex causal interrelationships typical of health care’.

However, a significant limitation of PRA is the lack of available probability data (Ward et al 2010), which means that data must be based on assumptions made by experts. Even experts with a healthcare safety background can find these assumptions difficult to quantify, as colleagues and I discovered when using expert elicitation methods to predict the effectiveness of interventions to prevent medication errors. Additional complex modelling was required to generate predictive data from care mapping and research literature (Karnon et al 2007).

More recent studies of HFMEATM projects have identified other significant methodological limitations. Habraken et al (2009) studied 13 projects in the Netherlands which were acknowledged to be successful, but reported that the process was time costly, the risk assessment process was difficult to carry out and there was a lack of guidance related to

‘the identification of failure mode causes and effective actions [that] might influence the quality of the outcomes of the analysis’.

There are other drawbacks to prospective hazard analysis methods. For example, Linken et al (2005) recorded over 250 person-hours spent on the analysis of a large-scale sterilisation process and Kessels-Harbraken et al (2009) found somewhat differing results between prospective and retrospective methods applied to the same hospital scenarios.

Nevertheless, the literature on healthcare resilience, which is concerned with the ability of an organisation or process to resist or recover from failure, and research on identifying gaps and discontinuities in care (Nemeth et al 2008a), would suggest that prospective hazard analysis has some value.

In the project reported in *Publication 3,* process mapping was used to track steps in care in the supported discharge process for people with chronic obstructive pulmonary disease, and during supported care for discharged patients in the community. Taking a view similar to Wreathall and Nemeth (2004) and Nemeth et al (2008b) - that process mapping requires inputs from people from the sharp end of care - both clinical staff and patients were interviewed and observed in order to identify possible failure modes and discontinuities of care. The observation-based care mapping approach used in our study maps was not dis-similar to that reported by Nemeth and colleagues (2011), who clearly demonstrate the value of the method in an Emergency Care setting when comparing flow chart mapping with a map based on observation.

I have not been able to identify a combination of methods for triangulating failure modes elsewhere in the literature that is similar to that in *Publication 3*. However, the comments made by Nemeth et al (2008b), that the differences of staff perceptions between ‘work as done’ and ‘work as imagined’ may be a limitation of the study. In their study of care in an Emergency Department, Nemeth and colleagues (2011) used trained clinical observers who were able to find the gaps in care, while we were only able to use an observer with research training but without clinical skills. Nevertheless, gaps were found and, if one takes a similar definition of resilience as did Smith et al (2013), nurses did use resilient behaviour to bridge gaps that had been unidentified by the organisation until the study was carried out.

Facing the challenge of ranking potential hazards without supporting data, the Delphi approach (Hasson et al 2000) was chosen to rank the potential hazards. This had two main benefits. The two-round Delphi was able to engage the interest of heath service staff and allowed each to participate equally. It also enabled a stepwise series of choices that could be recorded and it took each member of staff only a relatively short time to complete the ranking, reducing the time pressure.

The main challenges to the resilience of the supported discharge process were communication gaps and discontinuities, between patients and staff and between groups of staff across community and secondary care. In trying to identify solutions to these actual and latent failures, it proved difficult to engage staff in the final discussion group because of time pressures, a finding replicated by others (Ward et al 2010). Nevertheless, the project was able to identify solutions to three of the top-ranked potential failures and two other problems were resolved outside of the discussions.

In their work on the potential of a PRA toolkit for the NHS, Ward et al (2010) noted, among other issues,-

‘One of the difficulties of undertaking PRA in the NHS was process variability, where risks might be at one level during the day and at an entirely different level during the night’ (quite unlike most manufacturing and engineering industries). Variability also occurs at the micro and macro levels’

and that

‘A risk-based evaluation built upon a mapping of the processes is a useful approach to facilitating the level of discussion that might lead to a unified, or at least a less variable, process’.

Our study produced similar findings at an earlier stage of Prospective Hazard Analysis usage in the NHS.

Waring (2013) argues that by prospectively assessing hazards we fail to focus on the diverse resources of safety, and thus are in danger of replicating a managerial or technocratic approach to learning and change. Nevertheless it can also be argued that the principle purpose of undertaking a prospective hazard analysis in healthcare systems is not just to improve safety by finding latent vulnerabilities and producing fixes – it can also highlight resilience in the system. FMEATM helped to reveal the hidden resources of resilience in patients and nurses in the operation of the supported discharge scheme.

Moves to reduce the impact of gaps that were discovered, such as the breakdown of the system for patients calling for help by telephone, the variability of processes at night compared with day time in the flexibility of the patient-initiated admissions system and the improvements subsequently made, allowed the whole system to be more resilient under challenge. Although the system was less pressured than that described in the discussion on ‘going solid’ by Cook and Rasmussen (2005), there were elements of the system that were at the margins of the acceptable performance boundary, some of which were identified through patient interviews and which were not known to staff prior to the project. The inclusion of patient-provided information in a safety project can make a contribution to improving system resilience.

Nurses had also built resilient behaviour into their everyday practice in seeking ways through the complexity of nurse-led admissions in a system which was typified by doctor-led admissions. Smith et al (2013) have described these professional workarounds as resilient behaviours and, whereas I might have previously seen this as good, flexible professional practice, these individual workarounds did also appear to make the system safer and more flexible – and thus more resilient.

On reflection the study could have benefited from explicitly seeking to identify gaps in care. Cook et al (2000) have explored discontinuities in care as vulnerabilities that can be bridged by workers (just as some of the staff and patients did in our study). But to bridge a gap is not to eliminate it and while some bridges are robust and reliable, others are frail, brittle and easily undone by outside circumstances. As Wreathall (2006) and Jeffcott et al (2009) indicate, fixes that do not come to the direct attention of the organisation do not contribute as much to the resilience as they would if they were identified and acted on through a more structured process such as HFMEATM. Practitioners are generally effective in bridging gaps so that failures occur only relatively infrequently. But where an organisation is not, or is only partially, resilient, gaps remain an ever present threat to safety.

**4.3 Studies based on retrospective case note review (Publications 4, 5 and 6)**

As indicated in Section 3.1 above, all of these publications derive from a large project which comprised two parts. First, a study in nine hospitals of the comparative effectiveness of two methods of retrospective case note review:- explicit criterion based review and structured implicit review. Second, a study in twenty hospitals to explore the relationship between quality markers and proxy measures of outcomes.

*Publication 4* reports on the comparative reliability of explicit criterion based review versus structured implicit review methods. The study to explore the relationship between quality and safety markers and proxy measures of outcomes in twenty hospitals is reported in *Publication 5*. Using case note review data that was collected during the outcomes study, *Publication 6* draws on the results of the two studies to explore how structured, judgement based, review can be used to review safety and quality of care provided for people who die while in hospital.

***Reliability of case note review (Publication 4)***

Although there are methodological challenges in the use of retrospective case note review it still remains the main practical choice for hospital retrospective reviews of safety and quality of care (Zeegers et al 2007, Hogan et al 2012). While there are variations in how review methods are applied – for example, in some instances the use of data from trigger tools in electronic data bases, or using initial screening to identify cases for review based on a finding of a possible error marker (Landrigan 2010) – there are essentially two approaches.

In criterion based review a pre-selected set of clinical markers are used as indicators of the safety and quality of care provided. Criterion based review is particularly useful when screening the quality of care across a large number of clinical units. For instance, Caretrack in Australia uses retrospective review based on a total of 522 explicit criteria covering 22 common conditions (Runciman et al 2012).

Implicit case note review is the alternative method, but its use has been plagued by poor inter-rater reliability of the method when two or more reviewers assess safety and quality of care on the same patient record. Implicit review requires the clinically trained reviewer to make a judgement on the quality of care provided and it is at the point of judgement that variability affects the comparability of the reviews. For example, although Hayward and Hofer (2001) used a five point ‘preventability’ scale to assist with decision making in their study of whether deaths in hospital were preventable, the inter-rater reliability was relatively low (Intra Class Correlation (ICC) = 0.34).

Previous personal experience of the case note review literature (The North of England Study of Standards and Performance in General Practice 1992) informed decisions on the use of a phase of care structure for the implicit review. The structured review method reported here is a significant refinement over the criterion list/severity scale type of implicit review previously reported (Pearson et al 2000), since the method is enhanced by written comments that reviewers are asked to produce to support a judgement based score which they give for each phase of care and for care overall.

Case notes of people with chronic obstructive airways disease and with heart failure were reviewed in the nine hospitals. Physicians provided rather higher reliability scores (Intra Class Correlation (ICC) = 0.52) than those found by Hayward and Hofer (2001). Results suggested that nurse reviewers were more likely to comment on care pathways rather than on the safety of more direct interventions, though there was some indication that both aspects could judged by specialist nurses.

Weingart and colleagues (2002) used the analogy that

‘nurses couldn’t see the wood for the trees while physicians couldn’t see the trees for the wood’,

thereby implying that nurse reviewers in general take a more holistic approach but do not comment on the interventions whereas physicians take a more detailed interventionist approach, but may fail to see the wider healthcare picture. This seemed to be the case in the analysis presented in *Publication 4* although it later became clear that good physician reviewers are able to see both ‘wood and trees’ (*Publication 6*).

Additionally, work by Sharek et al (2010), using internal review teams to investigate patient harm, found that there was no significant difference between the results from internal and external reviewers, supporting our view that internal physician review teams are able to make appropriate care judgements about their own hospital.

***Quality assessment and hospital level outcomes (Publication 5)***

This phase of the study investigated potential relationships between the quality of care scores assessed by the study review methods (a more limited set of explicit review criteria together with structured implicit review) and a range of possible proxy outcome measures for hospitals. The study was undertaken on 1565 case notes in 20 randomly selected acute hospitals in England. For each of the index conditions a trained physician reviewer assessed up to 40 sets of case notes. Group training for the reviewers used the knowledge gained from the first phase of the study to ensure that judgement commentaries were as explicit as possible, rather than offering descriptive or implicit (and therefore obscure) information. Only the quantitative care scores were used in this publication, since there were not enough resources to analyse the extensive data set of commentaries.

The choice of proxy outcome measures drew on experience with measures reported in *Publication1* and *Publication 2*. In addition to those relatively few measures already shown to be associated with better quality and safety of care, the study included a number of markers that UK healthcare organisations were using or proposing.

Results reported in *Publication 4* indicated that the safety and quality scores from both the criterion based review and the structured implicit review were reasonably reliable and in *Publication 5* we were able to show that the scores could be used to rank hospitals. There were also some limited associations between the scores and safety climate items from the NHS staff survey. However, the analysis found no significant positive association between higher safety and quality scores and lower hospital mortality rates for COPD and for Heart Failure.

Nor was there any relationship between hospital level proxy outcome measures such as Hospital Standardised Mortality Ratios (HSMR) and the safety and quality scores. This was not altogether surprising since, contemporaneously, Pitches et al (2007) had shown in a meta-analysis that only approximately half of 51 correlations were positive in 36 process-outcome studies.

In *Publication 5* we saidthat

‘available hospital level outcome indicator data are probably insufficiently sensitive to reflect the quality of care recorded in patient case notes’.

Furthermore, we said that

‘The nuances of patient care may mean that high-quality,(safe) care may be given even when the patient’s outcome appears poor, and the reverse might also be the case. These (research) findings may be pointing to process measures as being more useful than outcome measures when reviewing the case notes of people who have chronic disease or multiple conditions’.

This may still be correct, although an alternative argument may be made that a greater number of case note reviews per condition in each hospital might have produced a different level of care score. Additionally, it is possible that the use of only two indicator conditions was insufficient to give an overall picture of the safety and quality of care for a hospital. Moreover, indices of severity of illness were unavailable, again limiting the strength of the analysis.

On the other hand, the commentary in *Publication 5* would have been more in favour of the strength of the case note review data if the qualitative analysis of the judgement commentaries had been available. As shown in *Publication 6* below, many of the 119 people who died had high overall care scores, and care scores in the study were usually coherent with the types of comments about the quality of care provided.

How may care scores contribute to a discussion on resilience? In an observational study of resilient actions in an oncology unit, Smith et al (2013) point out that the benefits of each of the actions they recorded tended to be for a particular patient and at a particular juncture of their care, without there being a general resilience benefit. Patterns of benefits or challenges are what is required to inform organisational and resilience improvements. Seen from a resilience perspective, therefore, the low phase of care scores might be used to screen for gaps in care (Nemeth et al 2008) or could be seen as markers that might indicate a brittleness of the system (Woods and Cook 2006). For example, if a group of reviews in the early admission phase found the scores were mainly low, or were low at times of the day or week, then questions could be asked about the service delivery system. These would need to be ‘behind the error’ questions (Woods et al 2010). For instance, Tucker and Edmondson (2003) took a ‘behind the error’ approach to hospital nursing practise and found that most process failures were not ‘errors’ but were ‘problems’ which were often neither recorded nor addressed by local leadership. Their research found that ‘problems’ happened close to once per hour of work per nurse, and so healthcare was unlike other industrial processes, where problems were infrequent. Problems were defined more broadly than errors, rather they were disruptions in a care givers ability to execute a prescribed task - discontinuities - and the authors argued that this material provides a rich learning opportunity. A similar case can be made for the judgement comments.

Exploring the value of phase scores further, where safety and quality scores are generally good (and over 60% of the cohort were shown to have good or very good care), the system might be seen as tending towards Hollnagel’s (2013) definition of safety –

‘the ability to succeed under varying conditions, so that the number of intended and acceptable outcomes is as high as possible’.

Nemeth et al (2011) describe acute care admission processes as

‘typically high-tempo, rapidly evolving, highly uncertain and involving high stakes’.

A set of good admission phase of care scores might reveal the presence of a safe and resilient team.

***Mortality review using structured judgement methods (Publication 6)***

During the outcomes study considered above, 1566 case note reviews were undertaken using structured judgement methods providing phase of care scores and associated judgement commentaries. Although an initial framework for a qualitative analysis of the judgement commentaries had been developed and was used to determine the differences between the types of responses made by physician and nurse reviewers (*Publication 5*), it had not previously been possible to explore the structured judgement comments further.

The increased NHS focus on understanding more about in-hospital deaths was a stimulus to explore whether the structured judgement method could contribute to a mortality review process. Three aspects of the policy and scientific debate around in-hospital death rates turned my attention to examining the care for the 119 people in our study who had died (7.6% of the total cases).

First, the Inquiry at the Mid Staffordshire NHS Foundation Trust (2013) had focused public and professional debate about safety of care and preventable deaths. Around the same time, research by Hogan et al (2010) had used a scale-score approach to explore the extent to which in-hospital deaths were preventable (around 5%). Hayward and Hofer (2001) had earlier found similar levels in the USA, although they raised questions about the interpretation of the term ‘preventability’ and suggested that, when predicted three month outcomes were taken into account, the proportion of preventable deaths was much lower than 5%.

Thirdly, there remains contentious debate over the use and value of statistical indicators used to compare mortality rates between hospitals, such as the Hospital Standardised Mortality Ratio (HSMR) (Speiglehalter 2013, Nicholl et al 2013). The HSMR was one of the indicators earlier used in *Publications 2* and *Publication 5*, and for which no associations were found with other safety and quality markers.

I therefore decided to explore the care for those who had died by examining the structured judgement comments for all (not just those with low scores who might have experienced poor care). In reviewer training the need to provide explicit rather than implicit judgements of care was emphasised, whether care was poor or good. Assessment of the over 700 judgement comments brought into focus how critically insightful some of the reviewers could be, for both technical and whole patient care. More than 80% of the judgement comments were explicit and most were associated with appropriate care safety and quality scores.

Analysis also highlighted that the structured judgement method has the advantage of being equally valuable in demonstrating experiences of good care, even when the outcome of the hospital episode is the death of a patient. Many of the people who died were found by reviewers to have had good technical care at the same time as having been provided with good holistic palliative care. Thus, this method can be used in a mortality review not just to review deaths where care was poor, or the death potentially avoidable, but also to celebrate excellent standards of care. The new ideas on safety also apply here –

‘the ability to succeed under varying conditions, so that the number of intended and acceptable outcomes is as high as possible’ (Hollnagel 2013) –

even if the overall outcome is a death that is not avoidable.

The strength of this mortality review method is not so much for the evaluation of an occasional high profile case but rather as a tool that allows questions to be asked of groups of cases - groups of patients within hospital units, or between similar units in different hospitals.

The literature on ‘gaps’ in care (Cook et al 2000, Nemeth et al 2008 a), and on the development of resilience markers (Furniss et al 2011b), sees trends in poor care as opportunities for improvement and to assist an organisation to be more proof to challenge. Most of the patients in our study were acutely ill with life threatening conditions. Most had arrived in hospital through an emergency department or an acute admissions unit, often out of hours, in situations that can become examples of an organisation operating outside of the safety envelope (Cook and Rasmussen 2005), certainly fertile ground for care that is ‘messy’ (Nemeth et al 2004).

Phase of care scores can act as markers for questions to be asked. If, for example, a number of reviews were found to be scoring low in the admissions phase, an exploration of the possible causes might be undertaken. Judgement comments are there to provide more detailed pointers. Phase of care review could also be seen as acting as a resilience marker, for example in a ‘de-compensated’ admissions unit setting. Intervention followed by further review might demonstrate change.

Some reviewers were able to identify critical trends such as lack of senior review at night or over weekends. Occasionally, a significant event is found. However, most poor care seemed to comprise constellations of smaller events (Hutchinson et al 2008). Alternatively, on a positive note, reviewers saw the value of a senior review when admitting teams were having difficulty in matching the right care plan to the diagnosis. So the phase of care quality scores can act as a screening tool and the judgement comments can act as an indicator.

How could resilience be enhanced? Findings from the reviews are not the place for a ‘root cause’ analysis but for a look ‘behind the error’ (Woods et al 2009). The phase of care framework might show that there are problems which on closer inspection are related to the (in)ability of the organisation to act quickly to prevent a decent towards ‘freefall’ (Wears et al 2006). Or patterns of poor care, when explored, might reveal that admissions policies are not given the required resources so that fixes, (Jephcott et al 2009, Smith et al 2013) applied by some resilient clinicians who are ‘taking it in their stride’ (Cook and Nemeth 2006), are not present when the ‘fixer’ clinicians are not available.

Improvements at the sharp end of the work will not be sufficient to affect resilience in the longer term. Clay-Williams (2013) points out that resilience, like safety (Weick 1987), is dynamic and is

‘characterised by persistence, adaptability, variability and unpredictability’.

Urgent action to correct a perceived problem should be restrained lest there are other complexities and vulnerabilities inadvertently introduced (Cook et al 2008). Rankin et al (2011) suggest

‘a shift in focus from human inability to human ability…is necessary for future work in safety management’.

The structured judgement review system should therefore be used to assist in understanding how healthcare requires change, for looking ‘behind the error’ – and not as a means of pointing up the need for targeted urgent action upon individuals. Complexity is likely to be an underlying characteristic of many of the failures (Woods et al 2007), including those identified by retrospective analyses of case notes. The challenge for review teams and the institutions where problems might have arisen is how to explore why the failure arose at the ‘second story’ level, without taking the ‘easy route’ of blame. Wears and Nemeth (2007) captured this approach with the title ‘replacing hindsight with insight’.

Phase of care structured judgement comment analysis may at least help to focus on the operational points where failures are occurring. Contributions to system resilience might therefore be made by the routine use of structured judgement review in a hospital, a view supported by Weick and Sutcliffe (2007) who see one aspect of a system’s resilience is its ability to detect ‘weak, surprise signals’ from routine performance data, that then triggers action by the organisation.

In their discussion on ‘organising for resilience’ Sutcliffe and Vogus (2003) argue that organisational resilience is not rare but arises

‘from relatively ordinary adaptive processes that promote competence, restore efficacy and encourage growth’.

But they also warn that these are not universal competencies in organisations. Moreover, as Nemeth and colleagues (2007) observe, there is a tension between clinical objectives for the individual patient and the management objectives across the organisation. To achieve improvement, a partnership approach between management and clinicians is therefore essential.

Management must therefore have a key role in improving safety and resilience. High level commitment from leadership (Edmondson 2004) is a prerequisite to improving performance. It should ensure, through the management structure and institutional policies, that the task of managing sick people has the right strategies and is supported by appropriate financial and staffing resources to achieve a resilient and safe care setting. Perhaps one of the most important additions to these resources will be investment in asking the deeper questions about failures within the service. While the structured judgement method can produce valuable data on care and safety quality, it is having the resources to understand the meaning of the results that will be critical in enhancing the safety and resilience of the organisation.

**5. Discussion**

The presented studies represent a body of methodological research intended to support improvements in safety and quality in health services. Although the work has had some international impact, the principal focus has been on UK health services. The challenge for the writer has been to take a retrospective view on a series of publications while at the same time seeking to put the findings in a fresh light - of a new perspective on safety and the emerging ideas on the meaning of resilience in healthcare.

I chose to use these new paradigms because I believe both the positive approach to safety espoused by Hollnagel and others (2013), and the ideas on building resilience into hard pressed health service settings, will eventually bring benefits to patients. At the time of writing there is only a limited UK literature related to ‘new safety’ and to healthcare resilience and this thesis has sought to point to the value of aspects of my work that can be seen as identifying and supporting positive and resilient behaviour in UK health services.

Other than the prospective hazard analysis study (Dean et al 2007) there are no direct examples in my work of resilient behaviour. Each of the other five publications are concerned with examples of what could be termed resilience markers, although these differ from the system-based markers proposed by Furniss et al (2011a, 2011b) because the publications refer either to safety climate dimensions or pointers such as phase of care scores.

However it is a feature of the literature on healthcare resilience as a whole that a fair proportion of the examples are not of research projects but rather are insightful case reports of system collapse and recovery in services, such as those in the USA of an Emergency Department (Wears et al 2006) or in a paediatric hospital (Cook and O’Connor 2005). This paucity of a research base is partly because of the relatively recent emergence of the science of healthcare resilience, particularly in UK health services research, but also of the methodological challenges of undertaking the work.

Embedding the results of resilience research into health services may be challenging. Much of the literature on the ‘new safety’ and resilience is insightful, sometimes elegant and often not particularly recent. But despite the efforts of authors referred to in this thesis, the use of markers such as judgement-supported phase of care scores may yet precipitate a ‘blame and train’ approach unless the new ideas on safety and resilience become embedded in health services. Perhaps the remarks from Berwick and colleagues on the need for a change in UK health service safety culture will help to turn the corner (Department of Health 2013).

***5.1 Limitations***

A methodological limitation of this thesis stems from the presentation of six publications that were written from a safety and quality perspective but not directly from a healthcare resilience viewpoint. All of the subject areas of the studies reported here are relevant in some aspects to the notion of resilience. But the retrospective approach of exploring a body of previously published research and setting it into a still emerging conceptual framework, sometimes related to studies from outside of healthcare, has proved challenging. Furthermore, a retrospective of methodological studies means that in some cases the methods, if not the messages, have been overtaken by more current literature.

So while there are methodological and conceptual linkages across all of the studies, there are limitations to the extent to which the studies make a direct contribution to resilience research. Nevertheless, the work seeks to present a view of where instances of resilience may lie in services and in the results of healthcare evaluation. Furthermore, the more recent publications on case note review and phase of care evaluation may offer a new approach to the timely identification of gaps by healthcare providers themselves, without the need to have specialist investigations, at least in the early stages of exploration.

**6. Reflection**

The studies from which the papers have derived were undertaken between 2002 and 2009. The first three publications resulted from interests that arose during an extended USA visit to a number of research teams engaged in the Agency for Healthcare Research and Quality (AHRQ) patient safety programme. This was a very fulfilling time – so many research ideas presented themselves – and it was one of the most exciting periods of my research career.

The research on case note review had a longer history for me. During an earlier career phase I had been involved in a large study of standards and performance in General Practice. The international literature from those times remained of interest to me and I was able to add this to the more recent literature on case note review for my recent research. Again, though, the 2002 research trip proved invaluable in understanding what was happening in the field, pre-publication, in the USA. And many of the research teams that I visited continued to be helpful and supportive during the fieldwork years.

Using the studies as a basis for the thesis provided both opportunities and challenges. With the emerging concepts of healthcare resilience and new perspectives on safety this was a good time to be able to reflect on some of the research I had undertaken over the preceding decade. Given the range of my publications round safety and quality as a topic, my research was carried out using a variety of methods, on different topics in differing settings. How was this to be managed within the context of the discussion in Section 4?

Formulating a coherent theme for this thesis was therefore the first challenge. The generic topic of safety and quality, which underpinned all of the work, appeared too broad and lacking in a focused concept. Likewise, concentrating on the multi-methods approaches was, in itself, not sufficient either for the reader or for me.

In developing the conceptual model for the thesis I first considered whether the notion of ‘risk’ was the concept I wished to use to link the publications and I explored this for some time. But, while acknowledging that all of the studies were about identifying risk in some way, I felt that I wanted to consider the work in a more anticipatory framework. At this point my interest in the concepts of resilience and resilience engineering were re-awakened.

As the topic of resilience in healthcare has been explored more fully in the literature over the past ten years it is clear that there is much that is relevant to my own work, although none of my work was framed specifically as resilient. Revisiting the more recent work on resilience was interesting now that I had a writing aim in mind. Yet some of the earlier work on resilient leadership, mindfulness and sensemaking, which I had discovered during my 2002 study tour, also remained exciting for me.

I am not alone as a research student in finding that the richness of the literature is in itself a challenge, because of the need for focus and, in the context of this form of thesis, the need to contain the new writing within a strict word limit. This has not prevented me from exploring some of the literature more widely, and enjoying the journey.

**7. Implications for research and practice**

**7.1 Implications for Research**

The following research questions arise from the safety and resilience analysis in the thesis. There are, of course, many health services research opportunities in the three fields of safety culture, prospective hazard analysis and case note review-based safety and quality research. The four suggested here reflect both a personal interest in mixed-methods research and a real desire to explore issues that might in the longer run bring benefits to patient care.

* Safety culture in health care – do resilient behaviours differ between staff from institutions with higher and lower staff survey safety culture data. Can the safety climate results available for the NHS Staff Survey be used to identify potential study sites.
* Prospective Hazard Analysis – despite the challenge of staff input availability, can PHA have a place in assessing the resilience of new care pathways or systems at an early stage of implementation?

Case note review:-

* Can the results of structured judgement review have a place as a resilience marker in clinical services.
* When using structured judgement review, is the care profile on the safety and quality axis the same for people who survive serious illness as for those who die (and if so, why)?

**7.2 Implications for Practice**

Implementation research suggests that many of the important results concerning health care effectiveness will not become part of routine service provision. Reviewing my presented publications in the context of the international literature leads me to suggest three areas for service development

* Top and middle leadership (both managerial and clinical) have a key role in developing safer and more resilient organisations. It should be a priority for every NHS provider organisation to understand to what extent the management of the organisation meets these twin goals and seeks to improve from their current position.
* Services should be planned with resilience in mind. Despite the challenge of limited resources, the safety and the value of targeted resource redundancy should always be considered (for example, in constructing a safe Hospital at Night programme)
* Safety and quality of care review in provider units needs to move away from ‘error’ and to build on the evidence base from which this thesis draws, thus enabling providers to understand the circumstances behind failure and to build a resilient future.

**8. Conclusion**

The six publications presented in this thesis comprise some of my contribution to research in the field of healthcare safety. I have sought to see this work in the light of recent ideas on the value of human and organisational resilience in health care, and from the stance of the more positive view of healthcare safety.

I show that, even though the studies may use different assumptions and methods, there are aspects in each one which can contribute to the debate on the value and assessment of resilient practise.

And from the perspective of the recent discussions on ‘what is safety’ (Hollnagel 2013), which sees the term ‘safety’ as relating to the

‘ability to succeed under varying conditions, so that the number of intended and acceptable outcomes is as high as possible’,

a fresh look has helped to illuminate the work as not being focussed on success or failure. Rather the variation in care can be partly seen as a result of resilient humans working in a complex, sometimes brittle, environment and that some of the presented research methods can identify points in care where a systems approach is required, rather than focussing on individuals.

My aspiration is that the learning from this research will assist clinical leaders and health service managers in their work with frontline staff in a way that treats variability as a generally positive attribute – as a marker for improvement rather than a process that requires containment.

**9. References**

Agency for Healthcare Research and Quality: Quality Indicators. [www.qualityindicators.ahrq.gov/psi\_overview.htm](http://www.qualityindicators.ahrq.gov/psi_overview.htm) (accessed 20 November 2006).

Akselsson R, Ek Å, Koornneef F, Stewart S & Ward M. (2009) Resilience safety culture*. In Proceedings of the 17th World Congress on Ergonomics (IEA).* Beijing China, International Ergonomics Association.

Albuhairan B, McIntosh A, O’Hara R & Hutchinson A. (2008) Safety policy development in an acute hospital: a case study of Patient Safety WalkroundsTM. *In*: *Creating and designing the healthcare experience; Conference Proceedings* *2008*. Strasbourg, The International Ergonomics Association. p.1-5.

Alonazi MS, Hutchinson A, O'Hara R & Freeman JV. (2011) Cross cultural challenges in the use of patient safety questionnaires: a case study from Saudi Arabia. *In* Albolino S *et al*. (eds) Healthcare Systems Ergonomics and Patient Safety. London, Taylor and Francis Group.p.179-180. *ISBN 978-0-415-68413-2*

Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, Etchells E, Ghali WA, Herbert P, Majumdar SR, O’Beirne M, Palacios-Ferflinger L, Reid R, Sheps S & Tamblyn R. (2004) The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. *Canadian Medical Association Journal,* 170(11),1678-86.

Battles JB, Stevens DP. (2009) Adverse event reporting systems and safer health care. *Quality and Safety in Health Care*, 18, 2.

Bristol Royal Infirmary Inquiry (2001) The report of the public inquiry into children’s heart surgery at the Bristol Royal Infirmary 1984-1995. London: Stationary Office.

Benn J, Koutantji M, Wallace L, Spurgeon P, Rejman M, Healey A & Vincent C. (2009) Feedback from incident reporting: information and action to improve patient safety. *Quality and Safety in Health Care*,18,11-2.

Braithwaite J, Clay-Williams R, Negus P & Plumb J. (2013) Health Care as a complex adaptive system. *In* E Hollnagel *et al.* (eds) *Resilient Health Care*. Farnham, Ashgate Publishing. p 57-73.

BrennanTA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Newhouse JP, Weiler PC, & Hiatt HH. (1991) Incidence of adverse events and negligence in hospitalised patients: results of the Harvard Medical Practice Study 1. *New England Journal of Medicine,* 324, 370- 376.

Clay-Williams R.(2013) Restructuring and the resilient organisation: implications for health care. *In* E Hollnagel *et al.* (eds) *Resilient Health Care*. Farnham, Ashgate Publishing p123-133.

Codman, EA. (1920) *A Study in Hospital Efficiency. As Demonstrated by the Case Report of the First Five Years of a Private Hospital*. Boston: Thomas Todd Co.

Cohen MR, Senders J & Davis NM. (1994) Failure mode and effect analysis: a novel approach to avoiding dangerous medication errors and accidents. *Hospital Pharmacy*, 29, 319-324.

Cook RI, Render M & Woods DD. (2000) Gaps in the continuity of care and progress on patient safety. *British Medical Journal*, 320, 791-794.

Cook R, Rasmussen J. (2005) Going solid: a model of system dynamics and consequences for patient safety. *Quality and Safety in Health Care*, 14, 130-134.

Cook RI and O’Connor M. (2005) Thinking about accidents and systems. *In* H Manasse *et al.* (eds) *Improving medication safety*. Bethseda, MD, American Society for Health-System Pharmacists. p1-21.

Cook RI, Nemeth CP & Dekker S. (2008 a) What went wrong at the Beatson oncology centre. *in* Hollnagel E *et al*. (eds) *Resilience Engineering Perspectives, Volume 1: Remaining sensitive to the possibility of failure*. Farnham, Ashgate Publishing. p 225-235.

Cook RI, O’Connor M, Render M & Woods D (2008b). Operating at the sharp end: the human factors of complex technical work and its implications for patient safety. *in* Manuel BM *et al.* (eds) *Surgical Patient Safety: Essential information for surgeons in today’s environment.*. Chicago, American College of Surgeons p19-30.

Cook R. Resilience, the second story, and progress on patient safety. (2013) *In* E Hollnagel *et al.* (eds) *Resilient Health Care*. Farnham, Ashgate Publishing pp 19-26.

Cook RI, Nemeth C. (2006) Taking things in one’s stride: cognitive features of two resilient performances. *in* Hollnagel E *et al.* (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company p 205-222.

Department of Health (2000a) *An organisation with a memory: report of an expert group on learning from adverse events in the NHS chaired by the Chief Medical Officer.* London, Stationery Office.

Department of Health (2001) *Report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995: Learning from Bristol.* London, Stationery Office.

Department of Health (2000b). *The report of the inquiry into quality and practice within the National Health Service arising from the actions of Rodney Ledward*. London, Stationery Office.

DeRosier J, Stalhandske E, Bagian J & Nudell T. (2002)Using health care failure mode effect analysis. *Journal of Quality Improvement,* 28, 248-267.

Edmondson A C. (2004) Learning from failure in health care: frequent opportunities, pervasive barriers. *Quality and Safety in Health Care,* 13 (Supplement ll), ii3-ii9.

Eiff G. (1999) *Organisational safety culture. Proceedings of the Tenth International Symposium on Aviation Psychology.* Columbus Ohio, Department of Aviation. p. 1-14.

Flin R, Mearns R, O’Connor P & Bryden R. (2000) Measuring safety climate: identifying the common features. *Safety Science*, 34,177-192.

Flinn R. Erosion of managerial resilience: from VASA to NASA. (2006) *in* Hollnagel E *et al.* (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company p 223-234.

Frankel A, Graydon-Baker, Neppl C, Simmonds T, Gustafason M & Gandhi TK. (2003) Patient Safety Leadership WalkroundsTM. *Joint Commission Journal on Quality and Safety*, 29,16-26.

Furniss D, Back J, Blandford A, Hildebrandt M & Broberg H. (2011) A resilience markers framework for small teams. *Reliability Engineering and System Safety*, 96, 2-10.

Furniss D, Black J, Blandford A. 2011 Unwritten rules for safety and performance in an oncology day care unit: testing the resilience markers framework. Proceedings of the Fourth Resilience Engineering Symposium. p.1-7.

Gaba DM. (2003) Safety first: ensuring quality of care in the intensely productive environment – the HRO model. <http://www.apsf.org/newsletters/html/2003/spring/hromodel.htm> (accessed 12 August 2013)

Habraken MMP, van der Schaaf TW, Leistikow IP & Reijnders-Thijssen PMJ. (2009) Prospective risk analysis of health care process; A systematic evaluation of the use of HFMEATM in Dutch health care. *Ergonomic*, 52(7), 809-819.

Hasson F, Keeney S & McKenna H. (2000) Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing,* 32, 1008-1015

Healthcare Commission NHS Staff Survey. [www.healthcarecommission.org.uk/nationalfindings/surveys.cfm](http://www.healthcarecommission.org.uk/nationalfindings/surveys.cfm) (accessed 20 November 2006).

Helmreich R and Merritt A. (1998) *Culture at work in aviation and medicine*. Aldershot, Ashgate.

Hogan H, Healey F, Neale G, Thomson R, Vincent C, & Black N. (2012) Preventable deaths in care in English acute hospitals: a retrospective case record study. *BMJ Quality and Safety in Health Care*. doi:10.1136/bmjqs-2012-001159.

Hollnagel E. Resilience: the challenge of the unstable. (2006) *in* Hollnagel E *et al.* (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company. p. 9-17.

Hollnagel E. Making healthcare resilient: from Safety I to Safety II.(2013) *In* E Hollnagel *et al.* (eds) *Resilient Health Care*. Farnham, Ashgate Publishing. p. 3-17.

Hutchinson A, McIntosh A, Coster JE, Cooper KL, Bath PA, Walters SJ & Pearson M. (2008) When is an event an event? Data from the quality and safety continuum. In: Hignett S *et al*. *From safe design to safe practice.* Cambridge: The Ergonomics Society. p. 179-183. ISBN 978-0-9554225-2-2.

Institute of Medicine. (2000) *To err is human: building a safer health system*. Kohn LT *et al* , (eds). Washington, DC: National Academy Press.

Jeffcott SA, Ibrahim JE & Cameron PA. (2009) Resilience in health care and clinical handover. *Quality and Safety in Health Care*,18, 256-260.

Kahn KL, Rubenstein LV, Sherwood MJ & Brook RH. (1989) *Structured Implicit Review for Physician Implicit Measurement of Quality of Care: Development of the Form and Guidelines for Its Use.* California, Rand Corporation.

Karnon J, McIntosh A, Dean JE, Bath PA, Hutchinson A, Oakley J, Thomas N, Pratt P, Freeman-Parry L, Karsh B-T, Gandhi T & Tappenden P. (2007) A prospective hazard and improvement analytic approach to predicting the effectiveness of medication error interventions. *Safety Science,* 45, 523-539.

Kessels-Habraken M, Van Der Schaff T, De Jonge J, Rutte C & Kerkvliet K. (2009) Integration of prospective and retrospective methods for risk analysis in hospitals. *International Journal for Quality in Health Care*, 21,427-432.

Landrigan CP, Parry GP, Bones CB, Hackbarth AD, Goldman DA & Sharek PJ. (2010) Temporal trends in rates of patient harm resulting from medical care. *New England Journal of Medicine*, 362, 2124-2134.

Linken DR, Sausman C, Santos L, Lyons C, Fox C, Saumiller L, Esterhai J, Pittman B & Lautenbach E. (2005) Applicability of healthcare failure mode and effects analysis to healthcare epidemiology: evaluation of the sterilisation and use of surgical instruments. *Clinical Infectious Diseases,* 4,1014-1019.

McNallyKM, Page MA & Sunderland BV. (1997) Failure-mode and effects analysis in improving a drug distribution system. *American Journal for Health-System Pharmacists*, 54, 171-177.

Nemeth CP, Nunnally M, O’Connor MF, Brandwijk M, Kowalsky J & Cook RI. (2007) Regularly irregular: how groups reconcile cross-cutting agendas and demand in healthcare. *Cognition, Technology and Work,* doi:10.1007/s10111-006-0058-4.

Nemeth C, Wears R, Woods D, Hollnagel E, Cook R. (2008a). Minding the gaps: Creating resilience in healthcare. *in* K Henriksen (eds) *Advances in Patient Safety: New directions and alternative approaches.* Vol. 3. Performance and tools. AHRQ Publication No. 08-0034-3. Rockville, MD, Agency for Healthcare Research and Quality. p. 259-271.

Nemeth CP, Cook RI & Woods DD. (2004) The messy details: insights from the study of technical work in healthcare. *IEE Transactions on systems, man and cybernetics. Part A: systems and humans*, 34, 689-692.

Nemeth C, Walter J, Wears R & Cook RI. (2008b) The path to resilience in ambulatory care. <http://www.ctlab.org/documents/AHRQ-PathToResillInAmbCare-Sep08.pdf> *Accessed 22/01/14.*

Nemeth C, Wears RL & Cook R. (2011) Resilience is not control: healthcare, crisis management, and ICT. *Cognition, Technology and Work*, doi 10.1007/s10111-011-0174-7.

Nicholl J, Jaques R & Campbell MJ. (2013) Mortality indicators used to rank hospital performance; should include deaths after discharge. *British Medical Journal*, 347, f5952.

Nieva VF, Sorra J. (2003) Safety culture assessment: a tool for improving patient safety in organisations. *Quality and Safety in Health Care,* 12, ii17-23.

O’Hara R, O’Keefe C, Mason S, Coster JE & Hutchinson A. (2010) Quality and safety of care provided by emergency care practitioners. *Emergency Medicine Journal*, doi: 10.1136/emj.2010.104190.

Patterson ES, Cook RI, Woods DD & Render ML. Gaps and resilience. *In* Bogner MS, (ed.) *Human error in medicine (2nd Edition)* Hilldale NJ, Lawrence Erlbaum Associates (*in press*).

<http://www.ctlab.org/documents/heinmedv26-gaps%20and%20resilience.pdf>

*accessed 22/1/14*

Pearson ML, Lee JL, Chang BL, Elliott M, Kahn KL & Rubenstein LV. (2000) Structured implicit review: a new method for monitoring nursing care quality. *Medical Care,* 38, 1074-1091.

Perrow C. (1999) *Normal accidents: living with high-risk technologies.* Princeton, NJ. Princeton University Press.

Pitches DW, Mohammed MA & Lilford RJ. (2007) What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Services Research*, 7, 91.

Rankin A, Lundberg J & Woltjer R. (2011) *Resilience strategies across industries for managing everyday risks. Fourth Resilience Engineering International Symposium.* Sophia Antipolis, France. URL. <http://urn.kb.se/resolve?urn=urn:se:liu:diva-80325> Accessed 20/01/2014

Reason, J. (1990) *Human Error.* New York, NY, Cambridge University Press.

Reason J. (1997) *Managing the risks of organisational accidents*. Farnham: Ashgate Publishing.

Reason J. (2000) Human Error: models and management. *British Medical Journal*, 320,768-770.

Reason J. Understanding adverse events: the human factor. (2001) *In* C Vincent. (ed) *Clinical risk management: Enhancing patient safety. 2nd Edition Ed*. London, BMJ Publishing Group, ISBN 0-7279-1392-1. p. 9-30.

Reason J. (2008). *The human contribution.* Farnham, Ashgate Publishing.

Rochlin GI, La Porte T & Roberts KH. (1987) The self designing high reliability organisation: aircraft carrier flight operations at sea. *Naval War College Review,* 42,76-90.

Roberts KH, Madsen P, Desai V & Van Stralen D. (2005) A case of the birth and death of a high reliability healthcare organisation. *Quality and Safety in Health Care,* 14, 216-220.

Rubenstein LV, Kahn KL, Harrison ER, Sherwood MJ, Rogers WH & Brook RH. (1991) *Structured implicit review of the medical record: a method for measuring quality of in-hospital medical care and a summary of quality changes following implementation of the Medicare prospective payment system*. Santa Monica CA, Rand.

Runciman WB, Hunt Td, Hannaford NA, Hibbert PD, Westbrook JI, Coiera EW, Day RO, Hindmarsh DM, McGlynn EA & BraithwaiteJ. (2012) Caretrack: assessing the appropriateness of health care delivery in Australia. *Medical Journal of Australia*, 197,100-105.

Sharek PJ, Parry G, Goldmann D, Bones K, Hackbarth A, Resar R, Griffin FA, Rhoda D, Murphy C, Landrigan CP. (2010) Performance characteristics of a methodology to quantify adverse events over time in hospitalized patients. Health Services Research, 654-678

Sexton JB, Thomas EJ & Helmreich RL. (2003) *Frontline assessments of healthcare culture: Safety Attitudes Questionnaire norms and psychometric properties*. *Technical Report 04-01*. Houston TX, The University of Texas Center of Excellence for Patient Safety Research and Practice.

Shapiro MJ & Jay GD. (2003) High reliability organisational change for hospitals: translating tenets for medical professionals. *Quality and Safety in Health Care*, 12, 238-239.

Shojania KG. (2008) The frustrating case of incident-reporting systems. *Quality and Safety in Health Care,* 17, 400-402.

Singer SJ, Gaba DM, Geppert JJ, Sinaiko AD, Howard SK & Park KC. (2003) The culture of safety: results of an organization-wide survey in 15 California hospitals. *Quality and Safety in Health Care*, 12, 112-118.

Smith MW, Giardina TD, Murphy DR, Laxmisan A & Singh H. (2013) Resilient actions in the diagnostic process and system performance. *BMJ Quality and Safety,* Published Online First: 27 06 2013. doi:10.1136/bmjqs-2012-001661.

Spiegelhalter D. (2013) Have there been 13000 needless deaths in 14 NHS trusts. *British Medical Journal*, 347, f4893.

Sutcliffe KM & Vogus TJ. (2003) Organizing for resilience. *In* Cameron KS *et al.* (eds) *Positive Organizational Scholarship: Foundations of a new discipline*. San Francisco, Berrett-Koehler Publishers Inc. p. 94-110.

Tanner J Padley W, Kiernan M. Leaper D, Norrie P & Baggott R. (2013) A benchmark too far; findings from a national survey of surgical site infection surveillance. *Journal of Hospital Infection*, 83, 87-91.

The North of England Study of Standards and Performance in General Practice. (1992) Medical audit in general practice I: Effects on doctors' clinical behaviour for common childhood conditions. *British Medical Journal,* 304, 1480-1484.

The Stationary Office. (2013) Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry. London: The Stationary Office. ISBN 9780102981476.

The University of Texas Center of Excellence for Patient Safety Research and Practice. Available at

http://www.uth.tmc.edu/schools/med/imed/patient\_safety/survey&tools.htm (accessed 25 August 2005).

Tucker AL, Edmondson AC. (2003) Why hospitals don’t learn from failures: Organisational and psychological dynamics that inhibit system change. *California Management Review*, 42(2), 55-72.

Ward J, Clarkson J, Buckle P, Berman J, Lim R & Jun T. (2010) *Prospective Hazard Analysis: tailoring prospective methods to a healthcare context.* Cambridge, University of Cambridge.

Waring J. (2013) What Safety II might learn from the socio-cultural critique of Safety I. *In* Hollnagel E *et al.* (eds) *Resilient Health Care*. Farnham, Ashgate Publishing p 39-48.

Waterson P, Griffiths P, Stride C, Murphy J & Hignett S. (2009) Psychometric properties of the Hospital Survey on Patient Safety Culture: findings from the UK. *Quality and Safety in Health Care,* doi 10.1136/qshc.2008.031625.

Wears RL, Perry S & McFauls A. (2006) Free Fall – A case study in resilience, its degradation and recovery in an Emergency Department. *In* Hollnagel E *et al* (eds) *Proceedings of the Second Resilience Engineering Symposium.* Paris France, Presses des Mines. p. 325-332.

Wears RL & Nemeth CP. (2007) Replacing hindsight with insight: toward better understanding of diagnostic failures. *Annals of Emergency Medicine,* 49, 206-208.

Weick KE. (1987) Organisational culture as a source of high reliability. *California Management Review,* 29, 112-127.

Weick KE & Sutcliffe KM. (2003) Hospitals as cultures of entrapment: a re-analysis of the Bristol Royal Infirmary. *California Management Review*, 45 (2), 73-84.

Weick KE & Sutcliffe KM. (2007) *Managing the unexpected – resilient behaviour in an age of uncertainty*. 2nd Edition. San Francisco, John Wiley and Sons Inc.

Weingart SN, Davis RB, Palmer RH, Cahalane M, Hamel MB, Mukamal K, Phillips RS, Davies DT & Iezzoni LI. (2002) Discrepancies between explicit and implicit review: Physician and nurse assessments of complication and quality. *Health Services Research*, 32, 483- 498.

Wilson RM, Runciman WB, Gibberd RW, Harrison BT, Newby L & Hamilton JD. (1995) The Quality in Australian Health Care Study. *Medical Journal of Australia*, 163(9), 458-71.

Woods DD, Cook RI. (2002) Nine steps to move forward from error.*Cognition Technology & Work,* 4, 137-144.

Woods DD. Essential characteristics of resilience. (2006a) *In* Hollnagel E *et al.* (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company. p. 21-34.

Woods DD, Patterson ES & Cook RI. (2007) Behind human error: taming complexity to improve patient safety. *In* Carayon P (ed). *Handbook of human factors and ergonomics in patient safety.* Mahwah, NJ, Lawrence Erlbaum Associates Inc. p. 459-476.

Woods DD & Cook RI. (2006) Incidents –markers of resilience or brittleness? *In* Hollnagel E *et al*. (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company. p. 69-75.

Woods DD, Dekker S, Cook R, Johannesen L & Sarter N. (2010) Behind Human Error. Farnham, Ashgate Publishing Limited.

Wreathall J & Nemeth C. (2004) Assessing risk: the role of probabilistic risk assessment (PRA) in patient safety improvement. *Quality and Safety in Health Care,* 13, 206-12.

Wreathall J. (2006) Properties of resilient organisations: an initial view. *In* Hollnagel E *et al*. (eds) *Resilience Engineering: Concepts and precepts.* Burlington VT, Ashgate Publishing Company. p275-285.

Zegers M, de Bruijne MC, Wagner C, Hounhoot LHF, Waaijman R, Smits M, Hout FAG, Zwaan L, Christiaans-Dingelhoff I, Timmermans DRM, Groenewegen PP & van der Wal G. (2009) Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. *Quality and Safety in Health Care*, 18, 297-302.

**10. Appendices to the Thesis**

10.1 Publication 1. Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability..........................................................57

10.2 Publication 2. Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System......................................................................................82

10.3 Publication 3. Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway............98

10.4 Publication 4. Assessing quality of care from hospital case notes: comparison of reliability of two methods......................................................117

10.5 Publication 5. Comparison of case note review methods for evaluating quality and safety in health care...................................................................118

10.6 Publication 6. A structured judgement method to enhance mortality case note review: development and evaluation....................................................336

**10.1 Publication 1**

**Hutchinson A**, Cooper KL, Dean JE, McIntosh A, Patterson M, Stride CB, Laurence BE, Smith CM.Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability.*Quality & Safety in Health Care* 2006;15:347-353

**Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability**

Hutchinson A, Cooper KL, Dean JE, McIntosh A, Patterson M, Stride CB, Laurence BE, Smith CM

**Section of Public Health, ScHARR**

University of Sheffield, Sheffield, UK

A Hutchinson

K L Cooper

J E Dean

A McIntosh

BE Laurence

CM Smith

**Institute of Work Psychology**

University of Sheffield, Sheffield, UK

M Patterson

C B Stride

**Corresponding author:** Professor Allen Hutchinson

Section of Public Health, ScHARR, Regent Court, 30 Regent St.

Sheffield S1 4DA, UK, Tel. 01142220811, Allen.Hutchinson@sheffield.ac.uk

Keywords: Safety climate, healthcare**Abstract:** **Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability**

**Aim**: The study aim was to explore the factor structure, reliability and potential usefulness of a patient safety climate questionnaire in UK healthcare.

**Setting**: Four acute hospital Trusts and nine Primary Care Trusts in England.

**Methods**: The questionnaire used was the 27 item “Teamwork and Safety Climate Survey”. Firstly, 33 healthcare staff commented on wording and relevance. Secondly, the questionnaire was sent to 3650 staff within the 13 NHS Trusts, seeking to achieve at least 600 responses as the basis for the factor analysis. 1307 questionnaires were returned (36% response). Factor analyses and reliability analyses were carried out on 897 responses from staff involved in direct patient care, to explore how consistently the questions measured the underlying constructs of safety climate and teamwork.

**Results**: Some questionnaire items related to multiple factors or did not relate strongly to any factor. 5 items were discarded. Two Teamwork factors were derived from the remaining 11 Teamwork items and three Safety Climate factors were derived from the remaining 11 Safety items. Internal consistency reliabilities were satisfactory to good, Cronbach’s alpha being 0.69 and above for all five factors.

**Conclusions**: This is one of the few studies to undertake a detailed evaluation of a patient safety climate questionnaire in UK healthcare and possibly the first to do so in primary as well as secondary care. The results indicate that a 22 item version of this safety climate questionnaire is useable as a research instrument in both settings, but also demonstrates a more general need for thorough validation of safety climate questionnaires before widespread usage.

**Using a safety climate questionnaire in UK healthcare: factor structure, reliability and usability**

As healthcare comes to be seen as a potentially high-risk environment, there is increasing pressure to assess the safety culture of health care organisations. Some authors promote the use of semi-structured or qualitative approaches to assessing culture [1] whilst others suggest that culture can be assessed using a questionnaire approach. [2][3]

There is a real debate over how effectively culture can be assessed using climate questionnaires. As described by Denison [4], culture refers to the ‘deep structure of organisations’ (values, beliefs and behaviours) and is traditionally assessed through interviews or observation, whereas climate mainly concerns individuals’ perceptions of their work environment (policies, practices and procedures) at a particular point in time [5][6][7][8] and is amenable to measurement by questionnaire [9]. Despite this debate, safety climate questionnaires have been used to assess safety culture in many safety-critical industries such as nuclear power and aviation [10] and petrochemicals [11]. Some questionnaires for use in healthcare have been derived from work in other industries such as aviation [12] or have included several items validated in other settings; for example, the armed forces [13]. Most of the available healthcare safety climate questionnaires have been developed in the United States.

It has been suggested that safety climate questionnaire data may be used as an indicator of aspects of the underlying safety culture [14] whilst others have promoted use of this data to evaluate safety programmes or track changes over time in a healthcare setting [15][16]. Now a number of US-developed safety climate questionnaires can be accessed via the internet [1][16][17] and are beginning to be used by US and UK healthcare organisations.

Although there is a slowly building research literature on the scientific basis for measuring safety culture in healthcare, there has been a warning that ‘the enthusiasm for measuring culture may be outpacing the science’ [2]. This may particularly be the case where there has been limited testing of how consistently the questionnaires measure specific domains of safety and also of whether questionnaires developed in one environment are valid for use in another healthcare setting. Such environmental differences might be at the national level (e.g. whether a US-developed questionnaire has the same meaning in a UK setting) or at the organisational level (e.g. whether a questionnaire designed for acute hospital care can also be used in primary care/general practice).

The aim of this study was to explore these questions by investigating the factor structure, internal reliability and usefulness of a US-developed safety climate questionnaire in a UK healthcare setting. An additional aim was to explore the relevance of the same questionnaire in both primary and secondary care. The study sought to achieve enough responses from each of the two settings to be able to undertake an analysis of the underlying factor structure of the questionnaire items (rather than attempting to assess the actual safety culture of the individual organisations, which would require a response rate of at least 50-60%).

**Methods**

*Selection of instrument*

A review of available patient safety climate questionnaires was undertaken in late 2003. Selection criteria were that 1) there was some public domain evidence base concerning development, 2) the instrument measured safety climate at the clinical team or directorate level, and 3) the instrument was short enough for use by busy health professionals. Using these criteria there was a limited selection from which to choose, some still being in final development stage [13][15][17] and others having limited peer reviewed data available, including the ‘family’ of the approximately 65 item “Safety Attitude Questionnaire” (SAQ) [2][12][18], the 19 item “Safety Climate Survey” [3][16][19] and the 27 item “Teamwork and Safety Climate Survey” [18].

Using our selection criteria we identified two of the shorter instruments (the Stanford University “Patient Safety Climate in Healthcare Organizations” questionnaire [13] and the “Teamwork and Safety Climate Survey” [18]) as possibly useful among frontline NHS staff.

Thirty-three healthcare professionals, 16 from primary care and 17 from acute hospital care, were asked to complete these two questionnaires and comment on their understanding of each item in the manner of a ‘thinking aloud’ protocol [20][21]. On the basis of this work, we concluded that the 27 item Teamwork and Safety Climate Survey contained a greater number of items that were applicable to front-line clinical teams; this instrument was therefore taken forward for further study. The original one-page format was retained and UK-related demographic questions added.

*The sample*

The purpose of using the questionnaire in a single-round study was to provide sufficient responses on which to carry out factor analysis and reliability analysis of the questionnaire structure. At least 300 responses were required for the factor analysis of each of the healthcare staff populations (acute care and primary care), a minimum of 600 in all. The sample size for the study was calculated accordingly, with the anticipation that the response rate may be as low as 20%.

Staff from 13 healthcare organisations were invited to take part. In the four acute hospital Trusts a random sample of 1900 recipients was drawn from staff involved in direct patient care, management, clinical support services or patient contact administration roles. The Teamwork and Safety Climate questionnaires were distributed through the hospital postal systems and returned in a reply-paid envelope direct to the study team.

In the nine Primary Care Trusts, the questionnaire was sent to 1750 staff, including General Practitioners, Practice Nurses and Practice Managers, and a sample of other staff (Community Nurses, Health Visitors, School Nurses and Allied Health Professionals).

*Analysis*

Factor analysis explores the extent to which individual items in a questionnaire can be grouped together according to the correlations between the responses to them, hence reducing the dimensionality of the data. If a questionnaire is to have construct validity, then the items should measure key underlying concepts (or factors) in a coherent way; items successfully measuring the same underlying factor should consistently generate similar responses to each other. The resulting groups of items can then be examined to interpret the meaning of the factors.

The questionnaire contained two sections, “Teamwork” and “Safety Climate” [18][25]. An initial exploratory factor analysis on all 27 items showed that the Teamwork questions factored out separately to the Safety Climate questions, but that there were multiple factors within each section. Exploratory factor analysis was therefore undertaken separately on each of the two domains of the questionnaire, Teamwork (14 items) and Safety Climate (13 items). This was carried out on a random 50% sample of respondents (the ‘construction’ half of the data), using principal components extraction. The number of factors to be extracted was determined using the Kaiser criterion (eigenvalues greater than 1) in conjunction with assessment of scree plots (the former method has a tendency to over-extract but in this case the two methods suggested identical solutions). Oblique rotation was used to aid interpretation of the resulting factor loadings. An optimal factor structure was derived and the internal consistency reliabilities of the resulting factors were assessed using Cronbach’s alpha. Analysis was performed in *M*plus using maximum likelihood estimation.

Confirmatory factor analysis was then undertaken to assess the fit of the proposed factor structure to the remaining 50% of the dataset. This was examined using a number of fit indices: CFI (comparative fit index), TLI (Tucker-Lewis index), RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual), in addition to the model chi square statistic. Missing items were listwise deleted, and items were treated as continuous variables.

To assess how well the factor model separately fitted each of the primary care and hospital datasets, and also whether there was a difference in the level of reported Teamwork and Safety Climate between the two groups, a test of factorial invariance and population homogeneity was carried out. This consisted of performing a series of confirmatory factor analyses, within which successive model parameters (inter-factor correlations, scale means, variances and factor loadings) were allowed to vary between the primary and secondary care groups. At each stage the results were examined to determine whether allowing the parameters to differ between the groups had improved the fit of the model (which would suggest a difference between the groups either in the factor structure or in the level of Teamwork / Safety Climate). Additionally, a confirmatory factor analysis was carried out on each of the primary and secondary care groups separately.

*Research governance*

External scientific review was provided by the Sheffield Health and Social Research Consortium. Ethical review was provided by North Sheffield Research Ethics Committee and research governance approval sought from each participating organisation.

**Results**

### Face validity

As a result of the ‘thinking aloud’ exercise, minor adaptations were made to the questionnaire wording before it was used in the survey. For example, ‘institution’ was changed to ‘organisation’ and ‘physicians’ to ‘doctors’. However, care was taken not to alter the underlying meaning of the items, and for this reason some wording was left unchanged; for example, the term ‘briefings’ (which was unfamiliar to a number of respondents) and ‘medical error’ (which several respondents associated only with doctors/medical interventions).

*Survey response rates*

Since the aim of the study was primarily to examine the factor structure, a single-round posting was used (with no follow-up requests to non-responders). 1307 responses were received, the overall response rate being 36% (33% for Primary Care Trusts and 38% for acute hospital Trusts). The number of responses and associated response rates were felt to be sufficient for factor analysis, since non-responders were unlikely to differ from respondents in terms of the pattern of relationship between their responses to different questions (i.e. the factor structure).

*Applicability of items to staff groups*

A number of responses were obtained from staff not involved in direct patient care, many of whom answered ‘not applicable’ to several questions. Factor analysis was therefore carried out only on responses from the 897 staff involved in direct patient care, which included 237 hospital nurses, 187 primary care nurses, 51 hospital doctors and 152 GPs.

Many questions, particularly in the Teamwork domain, exhibited very weak discrimination, with a high proportion of respondents answering ‘agree’ or ’strongly agree’. Item responses are shown in Tables 1 – 3 and the skew of the distributions in Table 4.

*Exploratory factor analysis*

Exploratory factor analysis initially suggested a three-factor solution for the Teamwork domain and a three-factor solution for the Safety Climate domain. However, the four negatively worded items consistently formed a separate factor. This occurred when analysing the Teamwork and Safety Climate domains separately and when analysing all 27 items together (Supplementary Table 7). Other authors [22][23] have suggested that artificial factors can occur as a result of respondents misreading negative items. Some of the staff we spoke to also commented that they had almost misread these items, especially since only a small proportion of items were reverse-worded. These items were therefore removed from the analysis at this stage (Table 3).

Exploratory factor analysis was then carried out on the remaining items in each of the Teamwork and Safety Climate domains. The optimal solution for the Teamwork domain contained two factors, which together explained 50% of the variance of the 11 items (Table 1). This suggested two underlying themes, which were interpreted as representing: 1) input into decisions and collaboration with other staff, and 2) information handover.

The best solution for the Safety Climate domain gave three factors explaining 61% of the variance (Table 2). From their item loadings, these were interpreted as representing: 1) attitudes to safety within own team and capacity to learn from errors, 2) overall confidence in safety of the organisation, and 3) perceptions of management’s attitudes to safety.

The factor loadings (a measure of how strongly each item relates to each factor) are shown in Tables 5 and 6.

*Internal consistency reliabilities*

Internal consistency reliabilities (how clearly a set of items measure a single theme) were satisfactory to good, with Cronbach’s alpha 0.69 or above in all five factors (Tables 1 and 2). Removing a further item from the initial 5 items forming Teamwork factor 2 improved the internal consistency reliability of this factor. This item (“Briefing staff on handovers between shifts is important for patient safety”) appears to relate to opinion, rather than what actually happens.

*Confirmatory factor analysis*

Confirmatory factor analysis on the remaining 50% (the ‘validation half’) of the dataset indicated an almost adequate fit of the model to the data, under the widely applied fit indices criteria [24]. The CFI and RMSEA took values of 0.93 and 0.08 for Teamwork, and 0.94 and 0.07 for Safety Climate respectively (Supplementary Table 8). These can be compared to the cut-off values for a good model, estimated as a CFI greater than 0.95 and an RMSEA smaller than 0.06 [24].

*Use of the questionnaire in different settings*

To assess how well the factor model fitted each of the primary care and hospital datasets, and also to explore whether there was a difference in Teamwork or Safety Climate between the two groups, a series of confirmatory factor analyses were performed, allowing successive model parameters to vary between the groups. Separate confirmatory factor analyses were also carried out on each of the primary and secondary care groups (Supplementary Table 9).

There was some difference in reported Teamwork Climate between the groups; the factor model showed a slight but significant improvement when scale means were allowed to vary, and mean scores on both Teamwork factors were lower for secondary care than for primary care. There also appeared to be some difference in optimal factor structure between the groups, since allowing the factor loadings to vary resulted in a significant improvement in fit (difference in χ2 = 39 on 9df, p < 0.05). Factor loadings for two items (“It is easy for staff here to ask questions when there is something that they do not understand” and “I have the support I need from other staff to care for patients”) were relatively low for the primary care sample (0.713 and 0.562), reflected in the low percentage of variance explained for these items for the primary care sample (R2 = 0.249 and R2 = 0.195, compared to R2 = 0.411 and R2 = 0.512 for secondary care). Separate confirmatory factor analyses on the two datasets indicated that the model fitted the secondary care data substantially better than the primary care data (CFI = 0.938 and CFI = 0.858 respectively).

For Safety Climate, the results of the multigroup analyses suggested that the optimal factor model was similar across both primary and secondary care, since improvement in fit was not statistically significant when factor loadings were allowed to vary between the groups. The separate confirmatory factor analyses on the two datasets indicated that the model offered a better fit to the hospital data, though the difference was smaller than for the Teamwork Climate model. However, there was evidence of a difference in level of Safety Climate between the groups, with the model fit greatly improved by allowing variation in the factor means (difference in χ2 = 43 on 3 df, p < 0.05) and variances (difference in χ2 = 15 on 3 df, p < 0.05). The mean scores on each of the 3 Safety Climate factors were significantly lower for the secondary care subsample, and their variances were greater.

**Conclusions**

The purpose of this study was two-fold: first, to test whether the questionnaire met conventional scientific criteria for internal reliability and factor structure and, second, to determine whether there was any difference in the factor model when compared between primary care and secondary care responses.

Factor structure

One key consideration when using questionnaires to assess safety climate is the need for a well-grounded item content (reflecting the topics to be covered by the questionnaire), together with a clearly defined factor structure relating groups of items (questions) to specific themes.

Interpretation of the factor structure of this questionnaire was not clear-cut, since some items were found to relate to more than one factor. This may cause difficulties in interpreting results if the questionnaire was to be used to evaluate safety programmes or track changes over time [15]. Conversely, the factors within a topic such as safety climate are likely to be inter-related to some extent, which may partially account for the cross-loading of several items to more than one factor (Tables 5 and 6).

Removing 5 items from the questionnaire improved the internal reliabilities of both domains. The three-factor Safety Climate solution was the more satisfactory, explaining 61% of the variance of the 11 items, whereas the two-factor Teamwork scale only explained 50%. Furthermore, our Teamwork factor 1 and Safety Climate factor 1 correspond reasonably well with the ‘Teamwork Climate’ and ‘Safety Climate’ factors in the 60-item Safety Attitudes Questionnaire from which the Teamwork and Safety Climate Survey questionnaire was originally developed [25] and which has recently been proposed for hospital-wide use [2]. Interestingly, our three Safety Climate factors also agree well with three key safety climate dimensions identified in reviews by Flin et al [14] and Wiegmann et al [26]. These relate to 1) employees’ own attitudes to risk and safety, 2) organisational commitment and safety system, and 3) management’s attitudes to safety.

Negatively worded items

In common with many other survey instruments, the Teamwork and Safety Climate questionnaire contained some items that were negatively worded (Table 3). Although suggested as a means of reducing response bias [27], negatively-worded items have often been found to factor out separately [22] [28] [29]. Schmitt and Stults [23] showed that an artificial factor can be produced when as few as 10% of respondents fail to recognise the reversal of the wording. Since the four negatively-worded items exhibited this clustering behaviour, they were removed from the final stage of the factor analysis.

*Use of the questionnaire in primary and secondary care*

We have also been able to show, we believe for the first time, that a safety climate questionnaire can be used across a whole health community, both in primary and in secondary care. The final overall factor model appeared to fit both groups reasonably well, although the fit was slightly better for the hospital data than for primary care, which is not surprising given that the questionnaire was designed for a hospital setting. This suggests that after further refinement (our interview data suggests that some questions may not be as relevant for primary care) comparisons using this type of questionnaire could be made between primary and secondary care.

However, a further cautionary note is that staff whose main role was not direct patient care answered ‘not applicable’ to many questions, which has implications for the use of this type of questionnaire across whole healthcare organisations.

*Limitations of the study*

The aim of this study was to evaluate the factor structure, and hence usefulness, of this safety climate questionnaire in UK healthcare. However, we did not aim to assess from first principles the key dimensions of safety climate in an NHS setting. This would have been the best strategy had we been setting out to create a new questionnaire but it would have been much more resource-intensive.

At this stage of the analysis we did not set out to undertake a multilevel confirmatory factor analysis to explore the influence that caregiver type (for example, doctor or nurse) might have had on the results.

*Implications*

The refined 22 item questionnaire provides a measure of safety climate in both primary and secondary UK healthcare, meeting some of the criteria on factor structure and internal reliability. It might, for example, prove useful in research studies that sought associations between safety culture and health outcomes, seeking also to determine the predictive validity of the instrument. However, there are enough cautionary points arising from the item content and factor analysis of this questionnaire to suggest that there is more to be done in exploring the properties of safety climate instruments, even those recently released [17], before proceeding wholesale into measuring safety climate across health services, at least those in the UK.

As Pronovost and Sexton [2] have recently pointed out, there is still much work required before we are able to understand the full value of using climate questionnaires in healthcare, including the meaningfulness of the resulting data. Until it is possible to derive evidence of predictive validity, such as whether positive culture data predicts measurably safer healthcare, the costs of routinely using safety climate questionnaires may not be justified.

**Acknowledgements**: We wish to thank the members of staff in all 13 sites who assisted with the organisation of the surveys. Our particular thanks are due to Eric Thomas and Bryan Sexton for their support in using the Teamwork and Safety Climate questionnaire.

Karen Beck acted as project administrator and her support was invaluable.

**Competing interests**: None declared by any of the authors

**Funding sources**: Funding was provided by the Sheffield Health and Social Care Research Consortium and by the devolved funds programme at the University of Sheffield.

**References**

1. National Patient Safety Agency “saferhealthcare” website. Available at

http://www.saferhealthcare.org.uk (accessed 25 August 2005).

1. PronovostP, Sexton B. Assessing safety culture: guidelines and recommendations. *Qual Saf Health Care* 2005;**14**:231-233.
2. Kho ME, Carbone JM, Lucas J, *et al*. Safety Climate Survey: reliability of results from a multicenter ICU survey. *Qual Saf Health Care* 2005;**14**:273-278.
3. Denison DR. What is the difference between organizational culture and organizational climate? A native’s point of view on a decade of paradigm wars. *Academy of Management Review* 1996;**21**:619-654.
4. Schneider B. The psychological life of organizations. In NM Ashkanasy, CPM Wilderon, MF Peterson (Eds.), *Handbook of organizational culture and climate*. Thousand Oaks, CA: Sage, 2000, xvii-xxi.
5. Barling J, Loughlin C, Kelloway EK. Development and test of a model linking safety-specific transformational leadership and occupational safety. *J Appl Psychol* 2002;**87**:488-496.
6. Griffin MA, Neal A. Perceptions of safety at work: a framework for linking safety climate to safety performance, knowledge, and motivation. *J Occup Health Psychol* 2000;**5**:347-358.
7. Zohar D. Safety climate: conceptual and measurement issues. In JC Quick, LE Tetrick (Eds.), *Handbook of occupational health psychology*. Washington, DC: American Psychological Association, 2003, 123-142.
8. Schein EH. Sense and nonsense about climate and culture. In NM Ashkanasy, CPM Wilderon, MF Peterson (Eds.), *Handbook of organizational culture and climate*. Thousand Oaks, CA: Sage, 2000, xxiii-xxx.
9. Helmreich R, Merritt A. *Culture at work in aviation and medicine*. Ashgate: Aldershot, 1998.
10. Mearns K, Flin R, Gordon R, *et al*. Measuring safety climate on offshore installations. *Work & Stress* 1998;**12**:238-254.
11. Sexton JB, Thomas EJ, Helmreich RL. Error, stress and teamwork in medicine and aviation: cross sectional surveys. *BMJ* 2000;**320**:745-749.
12. Singer SJ, Gaba DM, Geppert JJ, *et al*. The culture of safety: results of an organization-wide survey in 15 California hospitals. *Qual Saf Health Care* 2003;**12**:112-118.
13. Flin R, Mearns K, O’Connor P, *et al*. Measuring safety climate: identifying the common features. *Safety Science* 2000;**34**:177-192.
14. Nieva VF, Sorra J. Safety culture assessment: a tool for improving patient safety in healthcare organizations. *Qual Saf Health Care* 2003;**12**(Suppl ll):ii17-ii23.
15. Institute for Healthcare Improvement. Available at

http://www.ihi.org/IHI/Topics/PatientSafety/SafetyGeneral/Tools/Safety+Climate+Survey+%28IHI+Tool%29.htm (accessed 25 August 2005).

1. Hospital Survey on Patient Safety Culture. Agency for Healthcare Research and Quality, Rockville, MD. 2005. Available at http://www.ahrq.gov/qual/hospculture/ (accessed 25 August 2005).
2. The University of Texas Center of Excellence for Patient Safety Research and Practice. Available at

http://www.uth.tmc.edu/schools/med/imed/patient\_safety/survey&tools.htm (accessed 25 August 2005).

1. Sexton JB, Helmreich RL, Pronovost P, Thomas E. Safety Climate Survey. University of Texas. 2003.
2. Drennan J. Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *J Adv Nurs* 2003;**42**:57-63.
3. Ericsson KA, Simon HA. *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: MIT Press, 1993.
4. Roberts RE, Lewinsohn PM, Seeley JR. A brief measure of loneliness suitable for use with adolescents. *Psychological Reports* 1993:**72**:1379-1391.
5. Schmitt N, Stults DM. Factors defined by negatively keyed items: the result of careless respondents? *Applied* *Psychological Measurement* 1985;**9**:367-373.
6. Hu L, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling* 1999;**6**:1-55.
7. Sexton JB, Thomas EJ, Helmreich RL, *et al*. *Frontline assessments of healthcare culture: Safety Attitudes Questionnaire norms and psychometric properties*. Technical Report 04-01. Houston: The University of Texas Center of Excellence for Patient Safety Research and Practice, 2003. Available at http://www.uth.tmc.edu/schools/med/imed/patient\_safety/survey&tools.htm (accessed 25 August 2005).
8. Weigmann DA, Zhang H, von Thaden T, *et al*. *A synthesis of safety culture and safety climate research.* Savoy, Illinois: University of Illinois at Urbana-Champaign, 2002.
9. Nunnally JC, Bernstein I. *Psychometric Theory*. New York: McGraw-Hill, 1994.
10. Cordery JL, Sevastos PP. Responses to the original and revised Job Diagnostic Survey: is education a factor in responses to negatively worded items? *J App Psychology* 1993;**78**:141-143.
11. Spector PE, Van Katwyk PT, Brannick MT, *et al*. When two factors don’t reflect two constructs: how item characteristics can produce artifactual factors. *Journal of Management* 1997;**23**:659-677.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Teamwork factors and % responses to items** | | | | | | |
| **Teamwork factor 1:** Input into decisions and collaboration with other staff  Cronbach’s α = 0.84 | †Disagree strongly | †Disagree slightly | †Neutral | †Agree slightly | †Agree strongly | ‡Not applicable |
| Nurse input is well received where I work. | 1 | 3 | 6 | 23 | 66 | 4 |
| Decision-making where I work uses input from relevant staff. | 3 | 4 | 9 | 29 | 54 | 2 |
| The doctors and nurses here work together as a well co-ordinated team. | 3 | 8 | 9 | 33 | 47 | 4 |  |
| Disagreements where I work are resolved appropriately (i.e. not who is right, but what is best for the patient). | 4 | 7 | 16 | 35 | 38 | 2 |
| It is easy for staff here to ask questions when there is something that they do not understand. | 1 | 3 | 5 | 24 | 67 | 1 |
| I have the support I need from other staff to care for patients. | 1 | 5 | 5 | 27 | 62 | 1 |
| I am satisfied with the quality of collaboration that I experience with senior doctors where I work. | 5 | 9 | 12 | 35 | 40 | 11 |
| **Teamwork factor 2:** Information handover  Cronbach’s α = 0.69 |  |  |  |  |  |  |
| I know the first and last names of all the staff I worked with during my last shift / period of work. | 4 | 5 | 3 | 14 | 74 | 3 |
| Important issues are well communicated at shift changes / between periods of work. | 3 | 7 | 11 | 37 | 42 | 11 |
| Briefings are common where I work. | 4 | 7 | 13 | 34 | 41 | 8 |
| I am satisfied with the quality of collaboration that I experience with nurses where I work. | 2 | 4 | 5 | 32 | 57 | 4 |
| † Percentage of valid responses (not including “not applicable”) from direct patient care staff  ‡ Percentage of total responses from direct patient care staff | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2: Safety Climate factors and % responses to items** | | | | | | |
| **Safety Climate factor 1:** Attitudes to safety within own team; capacity to learn from errors  Cronbach’s α = 0.73 | †Disagree strongly | †Disagree slightly | †Neutral | †Agree slightly | †Agree strongly | ‡Not applicable |
| I am encouraged by my colleagues to report any patient safety concerns I may have. | 1 | 4 | 7 | 28 | 60 | 1 |
| The culture where I work makes it easy to learn from the errors of others. | 4 | 7 | 21 | 35 | 32 | 3 |
| I receive appropriate feedback about my performance. | 8 | 13 | 17 | 34 | 28 | 1 |
| Medical errors are handled appropriately here. | 1 | 4 | 18 | 27 | 50 | 5 |
| I know the proper channels to which I should direct questions regarding patient safety. | 1 | 4 | 6 | 30 | 59 | 1 |
| **Safety Climate factor 2:** Overall confidence in safety of organisation  Cronbach’s α = 0.70 |  |  |  |  |  |  |
| The levels of staffing where I work are sufficient to handle the number of patients. | 22 | 23 | 11 | 25 | 19 | 2 |
| I would feel safe being treated as a patient in this service. | 5 | 10 | 11 | 27 | 47 | 1 |
| Management does not knowingly compromise the safety of patients. | 6 | 6 | 14 | 22 | 51 | 2 |
| **Safety Climate factor 3:** Perceptions of management’s attitudes to safety  Cronbach’s α = 0.78 |  |  |  |  |  |  |
| This organisation is doing more for patient safety now than it did one year ago. | 4 | 5 | 36 | 28 | 27 | 5 |  |
| Leadership is driving us to be a safety-centred organisation. | 4 | 7 | 32 | 34 | 23 | 3 |
| My suggestions about safety would be acted upon if I expressed them to management. | 5 | 6 | 16 | 38 | 35 | 2 |
| † Percentage of valid responses (not including “not applicable”) from direct patient care staff  ‡ Percentage of total responses from direct patient care staff | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 3: Responses (%) to items omitted from final factor analysis** | | | | | | |
| **Teamwork items** | †Disagree strongly | †Disagree slightly | †Neutral | †Agree slightly | †Agree strongly | ‡Not applicable |
| \*Where I work, it is difficult to speak up if I perceive a problem with patient care. | 6 | 12 | 4 | 23 | 55 | 2 |
| \*I am frequently unable to express disagreement with the senior clinical staff here. | 9 | 13 | 14 | 27 | 37 | 6 |
| Briefing staff on handovers between shifts / periods of work (i.e. to plan for possible contingencies) is important for patient safety. | 0.4 | 0.3 | 3 | 12 | 84 | 15 |
| **Safety climate items** |  |  |  |  |  |  |
| \*Staff frequently disregard rules or guidelines (e.g. hand-washing, treatment protocols / clinical pathways, etc.) that are established for the area where I work. | 7 | 10 | 8 | 22 | 53 | 2 |
| \*Where I work, it is difficult to discuss errors. | 4 | 11 | 9 | 30 | 46 | 1 |
| † Percentage of valid responses (not including “not applicable”) from direct patient care staff  ‡ Percentage of total responses from direct patient care staff  \*These items are negatively worded | | | | | | |

|  |  |  |
| --- | --- | --- |
| **Table 4: Mean score for all items in each factor (scale)** | | |
| **Factor** | **Mean score** | **95% CI** |
| **Teamwork factor 1**  Input into decisions and collaboration with other staff | 4.25 | 4.21 – 4.30 |
| **Teamwork factor 2**  Information handover | 4.24 | 4.19 – 4.29 |
| **Safety Climate factor 1**  Attitudes to safety within own team; capacity to learn from errors | 4.10 | 4.05 – 4.14 |
| **Safety Climate factor 2**  Overall confidence in safety of organisation | 3.68 | 3.61 – 3.75 |
| **Safety Climate factor 3**  Perceptions of management’s attitudes to safety | 3.75 | 3.69 – 3.81 |

|  |
| --- |
|  |

|  |  |  |
| --- | --- | --- |
| **Table 5: Optimum factor loadings of Teamwork items1, 2** | | |
| **Teamwork items** | **Factor loadings** | |
|  | **Teamwork factor 1**  Input into decisions and collaboration with other staff | **Teamwork factor 2**  Information handover |
| **Teamwork factor 1** |  |  |
| Nurse input is well received where I work. | .697 |  |
| Decision-making where I work uses input from relevant staff. | .675 |  |
| The doctors and nurses here work together as a well co-ordinated team. | .830 |  |  |
| Disagreements where I work are resolved appropriately (i.e. not who is right, but what is best for the patient). | .716 |  |
| It is easy for staff here to ask questions when there is something that they do not understand. | .439 | .365 |
| I have the support I need from other staff to care for patients. | .510 | .420 |
| I am satisfied with the quality of collaboration that I experience with senior doctors where I work. | .779 |  |
| **Teamwork factor 2** |  |  |
| I know the first and last names of all the staff I worked with during my last shift / period of work. |  | .615 |
| Important issues are well communicated at shift changes / between periods of work. |  | .654 |
| Briefings are common where I work. |  | .456 |
| I am satisfied with the quality of collaboration that I experience with nurses where I work. |  | .532 |
| 1 Pattern matrix generated using principal components extraction and oblique rotation  2 Factor loadings under .300 omitted for clarity | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 6: Optimum factor loadings of Safety Climate items1, 2** | | | |
| **Safety Climate items** | **Factor loadings** | | |
|  | **Safety Climate factor 1**  Attitudes to safety within own team; capacity to learn from errors | **Safety Climate factor 2**  Overall confidence in safety of organisation | **Safety Climate factor 3**  Perceptions of management’s attitudes to safety |
| **Safety Climate factor 1** |  |  |  |
| I am encouraged by my colleagues to report any patient safety concerns I may have. | .707 |  |  |
| The culture where I work makes it easy to learn from the errors of others. | .623 |  |  |
| I receive appropriate feedback about my performance. | .663 |  |  |  |
| Medical errors are handled appropriately here. | .807 |  |  |
| I know the proper channels to which I should direct questions regarding patient safety. | .592 |  |  |
| **Safety Climate factor 2** |  |  |  |
| The levels of staffing where I work are sufficient to handle the number of patients. |  | .905 |  |
| I would feel safe being treated as a patient in this service. | .445 | .554 |  |
| Management does not knowingly compromise the safety of patients. |  | .519 | -.411 |
| **Safety Climate factor 3** |  |  |  |
| This organisation is doing more for patient safety now than it did one year ago. |  |  | -.903 |
| Leadership is driving us to be a safety-centred organisation. |  |  | -.786 |
| My suggestions about safety would be acted upon if I expressed them to management. | .383 |  | -.442 |
| 1 Pattern matrix generated using principal components extraction and oblique rotation  2 Factor loadings under .300 omitted for clarity | | | |

**Supplementary Table 7: Factor loadings of all 27 items1, 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Factor 1 (TW)** | **Factor 2 (TW)** | **Factor 3 (SC)** | **Factor 4 (SC)** | **Factor 5 (neg)** |
| **Factor 1 (Teamwork)** |  |  |  |  |  |
| Nurse input is well received where I work. | .591 |  |  |  |  |
| Decision-making where I work uses input from relevant staff. | .632 |  |  |  |  |
| The doctors and nurses here work together as a well co-ordinated team. | .853 |  |  |  |  |
| Disagreements where I work are resolved appropriately (i.e. not who is right, but what is best for the patient). | .597 |  |  |  |  |
| I have the support I need from other staff to care for patients. | .419 | .386 |  |  |  |
| I am satisfied with the quality of collaboration that I experience with senior doctors where I work. | .786 |  |  |  |  |
| **Factor 2 (Teamwork)** |  |  |  |  |  |
| It is easy for staff here to ask questions when there is something that they do not understand. |  | .481 |  |  |  |
| I know the first and last names of all the staff I worked with during my last shift / period of work. |  | .546 |  |  |  |
| Important issues are well communicated at shift changes / between periods of work. |  | .626 |  |  |  |
| Briefing staff on handovers between shifts / periods of work (to plan for possible contingencies) is important for patient safety. |  | .618 |  |  |  |
| Briefings are common where I work. |  | .531 |  |  |  |
| I am satisfied with the quality of collaboration that I experience with nurses where I work. | .309 | .481 |  |  |  |
| **Factor 3 (Safety Climate)** |  |  |  |  |  |
| I am encouraged by my colleagues to report any patient safety concerns I may have. |  |  | -.386 |  |  |
| The culture where I work makes it easy to learn from the errors of others. |  |  | -.612 |  |  |
| I receive appropriate feedback about my performance. |  |  | -.500 |  |  |  |
| Medical errors are handled appropriately here. |  |  | -.754 |  |  |
| I know the proper channels to which I should direct questions regarding patient safety. |  |  | -.599 |  |  |
| **Factor 4 (Safety Climate)** |  |  |  |  |  |
| The levels of staffing where I work are sufficient to handle the number of patients. |  |  |  | -.617 |  |
| I would feel safe being treated as a patient in this service. |  |  | -.304 | -.534 |  |
| Management does not knowingly compromise the safety of patients. |  |  |  | -.694 |  |
| This organisation is doing more for patient safety now than it did one year ago. |  |  |  | -.794 |  |
| Leadership is driving us to be a safety-centred organisation. |  |  |  | -.712 |  |
| My suggestions about safety would be acted upon if I expressed them to management. |  |  |  | -.644 |  |
| **Factor 5 (negatively worded items)** |  |  |  |  |  |
| Where I work, it is difficult to speak up if I perceive a problem with patient care. (Teamwork item) |  |  |  |  | .610 |
| I am frequently unable to express disagreement with the senior clinical staff here. (Teamwork item) |  |  |  |  | .594 |
| Staff frequently disregard rules or guidelines (e.g. hand-washing, treatment protocols / clinical pathways, etc.) that are established for the area where I work. |  |  |  |  | .765 |
| Where I work, it is difficult to discuss errors. |  |  | -.364 |  | .543 |

1 Pattern matrix generated using principal components extraction and oblique rotation

2 Factor loadings under .300 omitted for clarity

**Supplementary Table 8: Confirmatory factor analysis1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Section** | **χ2** | **CFI** | **TLI** | **RMSEA** | **SRMR** |
| Teamwork | 132 on 43df | 0.926 | 0.906 | 0.083 | 0.050 |
| Safety Climate | 116 on 41df | 0.942 | 0.922 | 0.071 | 0.049 |

1 For comparison, cut-off values for a good model have been estimated as: CFI greater than 0.95,

TLI greater than 0.95, RMSEA smaller than 0.06, SRMR smaller than 0.08 [ref 22]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Supplementary Table 9: Analysis of hospital and primary care subgroups**  **Teamwork domain** | | | | | | | | | | | | | | | | | | | | | | | |
| **(i) Test of factorial invariance** | | | | | | | | | | | | | | | | | | | | | | | |
| **Parameters allowed to vary between hospital and primary care** | | | **χ2** | | | | | | **Difference in χ2** | | | **CFI** | | | | **TLI** | | **RMSEA** | | | | **SRMR** | |
| All parameters constrained | | | 288 on 109df | | | | | | - | | | 0.851 | | | | 0.850 | | 0.104 | | | | 0.175 | |
| Factor correlations allowed to vary by subgroup | | | 286 on 108df | | | | | | 2 on 1df | | | 0.852 | | | | 0.850 | | 0.105 | | | | 0.166 | |
| Factor correlations and means allowed to vary by subgroup | | | 269 on 106df | | | | | | 17 on 2df \* | | | 0.865 | | | | 0.860 | | 0.101 | | | | 0.141 | |
| Factor correlations, means and variances allowed to vary by subgroup | | | 260 on 104df | | | | | | 9 on 2df | | | 0.871 | | | | 0.863 | | 0.100 | | | | 0.090 | |
| Factor correlations, means, variances and loadings allowed to vary by subgroup (all parameters free) | | | 221 on 95df | | | | | | 39 on 9df \* | | | 0.895 | | | | 0.879 | | 0.094 | | | | 0.065 | |
| **(ii) Confirmatory factor analyses run separately on subgroups** | | | | | | | | | | | | | | | | | | | | | | | |
| **Subgroup** | **χ2** | | | | **CFI** | | | **TLI** | | | **RMSEA** | | | | **SRMR** | | | | |
| Primary care | 122 on 43df | | | | 0.858 | | | 0.818 | | | 0.116 | | | | 0.075 | | | | |
| Hospital | 83 on 43df | | | | 0.938 | | | 0.921 | | | 0.075 | | | | 0.051 | | | | |
|  | | | | | | | | | | | | | | | | | | | | | | | | |
| **(b) Safety Climate domain** | | | | | | | | | | | | | | | | | | | | | | | | |
| **(i) Test of factorial invariance** | | | | | | | | | | | | | | | | | | | | | | | | |
| **Parameters allowed to vary between hospital and primary care** | | | | **χ2** | | | **Difference in χ2** | | | | | | **CFI** | | | | **TLI** | | **RMSEA** | | | | **SRMR** | |
| All parameters constrained | | | | 253 on 107df | | | - | | | | | | 0.879 | | | | 0.875 | | 0.087 | | | | 0.126 | |
| Factor correlations allowed to vary by subgroup | | | | 243 on 104df | | | 10 on 3df \* | | | | | | 0.884 | | | | 0.878 | | 0.086 | | | | 0.140 | |
| Factor correlations and means allowed to vary by subgroup | | | | 200 on 101df | | | 43 on 3df \* | | | | | | 0.917 | | | | 0.910 | | 0.074 | | | | 0.103 | |
| Factor correlations, means and variances allowed to vary by subgroup | | | | 185 on 98df | | | 15 on 3df \* | | | | | | 0.928 | | | | 0.919 | | 0.070 | | | | 0.065 | |
| Factor correlations, means, variances and loadings allowed to vary by subgroup (all parameters free) | | | | 170 on 90df | | | 15 on 10df | | | | | | 0.934 | | | | 0.919 | | 0.070 | | | | 0.056 | |
| **(ii) Confirmatory factor analyses run separately on subgroups** | | | | | | | | | | | | | | | | | | | | | | | | |
| **Subgroup** | | **χ2** | | | | **CFI** | | | | **TLI** | | | | **RMSEA** | | | | | | | **SRMR** | | | |
| Primary care | | 84 on 41df | | | | 0.928 | | | | 0.903 | | | | 0.079 | | | | | | | 0.063 | | | |
| Hospital | | 64 on 41df | | | | 0.964 | | | | 0.951 | | | | 0.053 | | | | | | | 0.051 | | | |

\* Indicates a statistically significant improvement in model fit at p < 0.05 level

**10.2 Publication 2**

**Hutchinson A**, Young TA, Cooper KL, McIntosh A, Karnon JD, Scobie S, Thomson RG. Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. *Quality and Safety in Health Care* 2009;18:5-10

Doi:10.1136/qshc2007.022400

**Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System**

**Hutchinson A, Young TA, Cooper KL, McIntosh A, Karnon JD, Scobie S, Thomson RG**

Section of Public Health, ScHARR, University of Sheffield,

Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK

A Hutchinson, Professor in Public Health

TA Young, Lecturer in Medical Statistics

KL Cooper, Research Associate

A McIntosh, Senior Research Fellow

JD Karnon, Senior Research Fellow

National Patient Safety Agency, 4-8 Maple Street, London, W1T 5HD, UK

S Scobie, Head of Patient Safety Observatory, NPSA

RG Thomson, Director of Epidemiology and Research, NPSA

**Corresponding Author:**

Professor Allen Hutchinson

Section of Public Health, ScHARR, University of Sheffield,

Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK

Email: [allen.hutchinson@sheffield.ac.uk](mailto:allen.hutchinson@sheffield.ac.uk) Tel: 0114 222 0811**Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System**

**Background:** Internationally, there is increasing recognition of the need to collect and analyse data on patient safety incidents, to facilitate learning and develop solutions. The National Patient Safety Agency (NPSA) for England and Wales has been capturing incident data from acute hospitals since November 2003.

**Objectives:** This study analyses patterns in reporting of patient safety incidents from all acute hospitals in England to the NPSA National Reporting and Learning System, and explores the link between reporting rates, hospital characteristics, and other safety and quality datasets.

**Methods:** Reporting rates to the NPSA National Reporting and Learning System were analysed as trends over time, from the point at which each hospital became connected to the system. The relationships between reporting rates and other safety and quality datasets were assessed using correlation and regression analyses.

**Results:** Reporting rates increased steadily over the 18 months analysed. Higher reporting rates correlated with positive data on safety culture and incident reporting from the NHS Staff Survey, and with better risk management ratings from the NHS Litigation Authority. Hospitals with higher overall reporting rates had a lower proportion of their reports in the “slips, trips and falls” category, suggesting that these hospitals were reporting higher numbers of other types of incident. There was no apparent association between reporting rates and the following data: standardised mortality ratios, data from other safety-related reporting systems, hospital size, average patient age or length of stay.

**Conclusions:** Incident reporting rates from acute hospitals increase with time from connection to the national reporting system, and are positively correlated with independently defined measures of safety culture, higher reporting rates being associated with a more positive safety culture.

***Keywords:*** *incident reporting, patient safety, safety culture*

**What this study adds**

Several countries have begun to implement national reporting systems for patient safety incidents in healthcare, in order to facilitate large-scale analysis of incident data. The number of incidents reported by different hospitals varies widely. However, little is known about the relationship between incident reporting rates and other hospital-level data on safety and quality of care.

This study indicates that incident reporting rates from acute hospitals to a national reporting system increase with time from connection to the system, and that rates are positively correlated with measures of safety culture, higher reporting rates being associated with a more positive safety culture.

**Introduction**

Since the publication of the US Institute of Medicine report “To err is human”[1] and the UK Department of Health report “An organisation with a memory”,[2] there has been increasing recognition of the need for healthcare organisations to monitor and learn from patient safety incidents. Proposals on how to accomplish this have included the use of reporting systems, and over the last few years several countries have established national or system-wide reporting systems to facilitate large-scale monitoring and analysis of incident data.[3, 4, 5] The National Reporting and Learning System (NRLS) for England and in Wales, established by the National Patient Safety Agency,[6] was rolled out from late 2003 and has now received over one million reports, mainly from acute hospitals.[7]

Although there may be benefits to be gained from the establishment of large reporting systems,[8, 9] there are challenges that accompany their development, both at the individual reporting level and at the data handling and analysis level. Many incidents still go unreported, with doctors being less likely than nurses to report.[10, 11] Barriers to reporting include time constraints, lack of knowledge about how and what to report, fear of blame, lack of feedback, and a perceived lack of value in the reporting process.[11, 12, 13] There are also technical challenges, such as connecting the national system to the many different local systems within health organisations, and developing a consistent framework for categorising incidents.[3, 14, 15] Furthermore, once the information has been captured, large amounts of data must be analysed so that meaningful feedback can be produced.[3] The World Health Organisation Draft Guidelines for Adverse Event Reporting and Learning Systems[8] identify a range of analytical approaches, including correlations and trend and cluster analyses, that might be used to identify patterns and assess risks.

All acute hospitals in England are now beginning to report to the National Reporting and Learning System, having been progressively connected to the system over a period of about two years. However, even taking into account the differential timetables for connection, it is clear that some organisations are reporting higher numbers of incidents than others.[7]

The National Reporting and Learning System is not the only quality and safety system to which hospitals in England are asked to report. For example, all are expected to report cases of bacteraemia due to Methicillin Resistant Staphylococcus Aureus to the Department of Health,[16] and to report incidents related to medicines and equipment to the Medicines and Healthcare Products Regulatory Agency.[17]

In addition, there are other data related to quality and safety that may have a relationship to incident reporting. For example, the annual NHS Staff Survey[18] includes questions about safety culture and incident reporting, and the NHS Litigation Authority[19] conducts a risk management assessment within each hospital and awards a rating on the basis of this.

The NHS in England therefore has an increasingly rich dataset relating to safety and quality in healthcare. This study has two main aims. Firstly, it analyses the patterns and trends in reporting by acute hospitals in England to determine whether there are any defining features of those which have higher or lower reporting rates. Secondly, it explores the relationship between reporting rates and other data relating to patient safety and quality of care.

**Methods**

Hospitals included in the analysis

Anonymised, aggregated data on the number and type of incidents reported by the 173 acute hospitals in England to the National Reporting and Learning System (NRLS) between April 2004 and November 2005 were provided by the National Patient Safety Agency.[6] Of the 173 acute hospitals, 148 were included in the analysis as they had a permanent connection to the NRLS and had reported at least one incident during the period analysed (April 2004 to November 2005).

Analyses undertaken and their objectives

Initially, an analysis of patterns and trends in reporting was undertaken to explore whether higher reporting rates were associated with particular hospital characteristics or types of incident (Table 1). Secondly, the relationship between reporting rates and other hospital-level safety data was explored (Table 2).

*Table 1: Analysis of patterns and trends in reporting*

|  |  |
| --- | --- |
| **Analyses** | **Datasets used** |
| Trend in reporting rates over time | Number of reports to NRLS per month per hospital, provided by the NPSA |
| Relationship with type/severity of incidents reported | Number of reports to NRLS per month per hospital categorised by incident type and severity, provided by the NPSA |
| Relationship with hospital size and characteristics | Healthcare Commission definitions of hospital type and size.[20] Data on patient age, gender, length of stay, and proportions of emergency and day-case admissions for 2004-5, obtained from the Hospital Episodes Statistics (HES) website[21] |

*Table 2: Analysis of relationships between reporting rates and other safety data*

|  |  |
| --- | --- |
| **Analyses** | **Datasets used** |
| Relationship with safety culture data | Proportion of staff at each hospital giving positive responses to the 2004 and 2005 NHS Staff Survey questions on safety culture and incident reporting, obtained from the Healthcare Commission website[18] |
| Risk management ratings per hospital awarded by the NHS Litigation Authority, ranging from 0 (lowest standard) to 3 (highest standard); data for December 2005 were obtained from the NHS Litigation Authority website[19] |
| Relationship with reporting rates to other systems | Number of reports of Methicillin-Resistant Staphylococcus Aureus (MRSA) bacteraemias during 2004-5, obtained from the Department of Health website[16] |
| Number of reports about medical devices to the Medicines and Healthcare Products Regulatory Agency (MHRA)[17] during 2005, obtained from the MHRA via the NPSA |
| Relationship with outcome data | Hospital Standardised Mortality Ratios (HSMRs) for each hospital for 2004-5, obtained from the Dr Foster website[22] |
| Hospital-level data on Patient Safety Indicators, measures of safety and quality which can be calculated using HES data. These indicators were adapted by the Healthcare Commission and the NPSA from those developed by the Agency for Healthcare Research and Quality (AHRQ) in the US;[23] data were provided by the NPSA. |

Calculation of reporting rates

To account for hospital size and activity, reporting rates were expressed as the number of reports per 100 bed days per month, or per 100 consultant episodes per month. Data on bed days and consultant episodes were obtained from the Hospital Episode Statistics (HES) website.[21]

Calculation of trends in reporting rates

The moving average method was used to calculate the trend in reporting rates,[24] in order to minimise fluctuations caused by data being reported and processed in batches, both within hospitals and within the NRLS. In this method, the first data point is an average of months 1,2,3; the second point is an average of months 2,3,4; and so on. The data points were calculated from each hospital’s connection date onwards.

Correlation and regression analyses

Linear associations between reporting rates to the NRLS and other safety data were explored using Spearman’s correlation coefficient and ordinary least squares linear regression analyses, using STATA.[25] Statistical tests were two-sided and a significance level of p ≤ 0.05 was used for all analyses.

**Results**

Trends and patterns in reporting to the National Reporting and Learning System

Figure 1 shows the trend in mean reporting rates to the NRLS across 148 hospitals (per 100 bed days per month). To minimise the effect of hospitals becoming connected to the system at different times, the data points are calculated from each hospital’s connection date onwards (i.e. month 1 is an average of the first live month of all hospitals, irrespective of date of connection). Reporting rates were still increasing when hospitals had been connected to the system for 18 months (mean reporting rates showed a linear trend; p < 0.001). Trends in reporting rate were similar whether expressed per 100 bed days or per 100 consultant episodes (not shown); rates per 100 bed days were selected as a proxy for hospital size and activity, since bed days are a more discrete entity and were thought to better reflect the level of hospital activity.

**Figure 1: Trend in mean number of reports per 100 bed days, from date of connection to NRLS**



Hospitals with higher overall reporting rates had a lower percentage of reports in the “slips, trips and falls” incident category (correlation -0.21, 95% confidence interval (CI) -0.37 to -0.05, p = 0.015). Figure 2 illustrates that while, for the majority of hospitals, the rate of falls increases in line with other types of incident, a few hospitals appear to be reporting higher rates of non-fall incidents while the rate of falls remains relatively constant. There were no significant correlations with proportions of other incident types. Nor was there a correlation between overall reporting rate and the proportion of reports relating to incidents of differing levels of severity.

**Figure 2: Relationship between “slips, trips and falls” and other types of incident (per 1000 bed days)**



Relationship between reporting rates and safety culture data

The NHS Staff Survey is completed annually by staff in all NHS organisations in England. We analysed the relationship between reporting rates to the NRLS and the proportion of staff at each hospital giving positive responses to NHS Staff Survey 2004 and 2005 questions on safety culture and incident reporting.[18] Of the six questions on fairness and effectiveness of reporting in the 2004 survey, there were significant linear relationships between higher reporting rates and a higher proportion of positive responses to four of the questions (Table 3). Although the wording and response categories were somewhat dissimilar in the 2005 survey, there were significant correlations on the question on encouragement to report (regression coefficient 0.03, 95% CI 0.01 to 0.06, P = 0.009) and on the question on blaming and punishing for making errors (regression coefficient 0.03, 95% CI 0.01 to 0.06, p = 0.002). Of the eight questions on health and safety at work in the 2004 survey, there was a significant linear relationship with one question: whether staff have access to counselling services at work (regression coefficient 0.02, 95% CI 0.004 to 0.036, p = 0.016). This question was not included in the 2005 survey.

*Table 3: Linear regression coefficients for predicting reporting rates from 2004 Staff Survey responses*

|  |  |  |
| --- | --- | --- |
| **Questions on fairness and effectiveness of reporting** | **Regression coefficients**  **(95% CI)** | **p-values** |
| Knows how to report errors, near misses and incidents | 0.06 (-0.01 to 0.12) | 0.080 |
| Employer treats fairly staff involved in an error, near miss or incident | 0.03 (0.005 to 0.06) | 0.021\* |
| Employer encourages staff to report errors, near misses or incidents | 0.05 (0.02 to 0.09) | 0.004\* |
| Employer treats reports of errors, near misses or incidents confidentially | 0.03 (0.01 to 0.06) | 0.014\* |
| Employer does not blame or punish people who make errors | 0.03 (0.005 to 0.05) | 0.017\* |
| When errors are reported, employer takes action to ensure that they do not happen again | 0.02 (-0.01 to 0.04) | 0.145 |

\* Significant at p ≤ 0.05 level

The survey also included questions on the number of errors observed and reported by staff. The correlation between higher reporting rates and the proportion of staff having seen at least one error in the last month was not significant (2004 survey, regression coefficient 0.02, 95% CI -0.001 to 0.05, p = 0.058). However, there was a significant positive correlation between higher reporting rates and a higher proportion of staff having reported the last error/near miss they saw (2004 survey data, regression coefficient 0.04, 95% CI 0.01 to 0.06, p = 0.005, 2005 survey data, regression coefficient 0.04, 95% CI 0.01 to 0.07, p = 0.015).

There was an additional correlation in the 2005 data in which as the number of staff indicating that they had a moving and handling injury increased there was a decrease in the number of reported overall incidents per 100 bed days per year (correlation coefficient -0.046, 95% CI -0.090 to -0.002, p = 0.041). A similar non-significant trend was found in the 2004 data.

The NHS Litigation Authority (NHSLA)[19] conducts risk management assessments within hospitals. Organisations can choose whether to be assessed for Level 1, 2 or 3, with Level 3 being the highest (Table 4), and receive a financial incentive for achieving a higher Level. We explored the relationship between reporting rates and NHSLA risk management ratings. The 10 hospitals achieving Level 3 had significantly higher reporting rates than hospitals at Level 1 or 2 (Table 4), although there was no significant difference in reporting rates between hospitals at Levels 1 and 2.

*Table 4: Relationship between reporting rates and NHS Litigation Authority (NHSLA) risk management ratings*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NHSLA risk management rating** | **Interpretation of ratings** | **Number of hospitals** | **Mean reporting rate per 100 bed days (95% CI)** | **Significant differences in reporting rates** |
| Level 0 | Not achieved Level 1 | 1 | 3.89 (n/a) | * Between Levels 1 and 3: p = 0.026 * Between Levels 2 and 3: p = 0.037 |
| Level 1 | Effective risk management systems and processes established | 88 | 0.82  (0.68 to 0.96) |
| Level 2 | Systems described for Level 1 have been implemented | 49 | 0.84  (0.61 to 1.06) |
| Level 3 | Organisation is monitoring compliance with systems and acting on findings | 10 | 1.36  (1.06 to 1.66) |

n/a – not applicable

Relationship between reporting rates and other safety and quality data

The numbers of reports to the NRLS were compared to the numbers of reports to other safety-related reporting systems. There were no significant correlations with the number of reports to either the Methicillin-Resistant Staphylococcus Aureus (MRSA) bacteraemia surveillance system[16] (correlation -0.05, 95% CI -0.22 to 0.11, p = 0.511) or the Medicines and Healthcare Products Regulatory Agency medical device reporting system[17] (correlation 0.08, 95% CI -0.09 to 0.25, p = 0.355). However, it should be noted that the numbers of reports to these systems were low, with an average of 42 reports to the MRSA surveillance system per hospital per year (95% CI: 36 to 47) and 15 reports to the MHRA medical device reporting system (95% CI: 13 to 18).

The relationships between reporting rates to the NRLS and selected patient outcome measures were also assessed. No significant correlations were observed, either with Hospital Standardised Mortality Ratios [22] (correlation 0.02, 95% CI -0.16 to 0.19, p = 0.865), or with three Patient Safety Indicators (see Methods) relating to a) number of deaths in low-mortality Healthcare Resource Groups, b) number of decubitus ulcers, and c) number of post-operative sepsis cases.

In addition, we explored the relationship between reporting rates and hospital characteristics. There was no significant difference overall between the mean reporting rates from different sizes and types of acute hospital as defined by the Healthcare Commission,[20] e.g. large, medium, small, teaching or specialist. Neither were there any significant correlations between reporting rate and data on patient age, gender, length of stay, waiting list time, or proportions of emergency or day-case admissions.[21]

**Discussion**

Patterns of reporting

Acute hospitals are reporting increasing numbers of incidents to the National Reporting and Learning System even after 18 months’ connection to the system. It will be interesting to monitor whether rates continue to increase, particularly following the provision of more structured feedback to hospitals which began in May 2006. The use of bed days as a denominator for reporting rates seems to be an appropriate way of accounting for hospital size and activity, and we found that using consultant episodes gave very similar results (not shown). It is also interesting to note that, while slips, trips and falls still account for around 40% of incidents being reported,[26] some hospitals appear to be reporting higher numbers of other incident types. Falls are traditionally reported by nursing staff, but our results suggest an improvement in willingness or ability of NHS staff to report other types of incident in addition.

Limitations of this research

This analysis was undertaken at an early stage in the lifetime of the NRLS, and it is worth noting that reporting rates may be affected by factors such as the technicalities of transferring data from local systems to the national system. The method used to calculate reporting rates took account of variations in when hospitals started to report to the NRLS, but may not adjust for this fully.

The relationship between reporting rate and safety culture

Since incident reports to a national system are a relatively new data source, contextual analysis alongside other data is critical for interpretation. The significant correlations between reporting rates and Staff Survey responses over the years 2004 and 2005 suggest that staff perceptions of the culture of safety and reporting within their hospital influence the actual number of reports being made. There was also a correlation between high reporting rates and the hospitals scoring highest in the NHS Litigation Authority risk management assessment. The positive association of actual reporting rates with independent data on safety culture is an important finding, and supports the view that, currently, higher reporting rates indicate safer organisations.

The apparent lack of a relationship with Hospital Standardised Mortality Ratios and Patient Safety Indicators may reflect the difficulty in demonstrating the link between safe, good quality care and patient outcomes.[27] The lack of correlation with reporting rates to other reporting systems (such as the MHRA medical device reporting system and the MRSA bacteraemia reporting system) is likely to be due to the relatively low numbers of reports to these systems, or to other factors such as the type of hospital and severity of casemix. It should also be borne in mind that while measures of safety are often collected at hospital level, both the reporting rate and the safety culture have been shown to vary widely between different departments within a single hospital, and also between different sectors of staff.[10, 11, 28]

Implications

Our findings suggest that higher reporting rates may be related to a more supportive culture of safety and reporting, as indicated by the relationship between reporting rates and data from the NHS Staff Survey and the NHS Litigation Authority. Reporting rates also appear to be improving over time across acute hospitals in England. The NPSA is now providing hospitals with regular feedback reports to allow hospitals to benchmark data in comparison with other similar hospitals, as part of a strategy to improve the quantity and quality of data reported. This analysis has been undertaken using nationally available data sources at hospital level. However, individual hospitals are likely to have access to additional data sources, which will aid the interpretation and analysis of NRLS reporting rates. Hospital reporting rates are not published more widely, or used directly as measures of compliance with health care standards. On the basis of our analysis, and in the context of making data available to support choice of care providers, we suggest that high reporting rates may be indicators of a positive safety culture, rather than markers of less safe care.

**Acknowledgements:** We wish to thank Patricia Bain and the other NPSA Patient Safety Managers who assisted with the study, and the MHRA for providing data.

**Contributorship:** AH wrote the research proposal, directed the study and contributed to analysis and writing, TY undertook much of the analysis and contributed to the writing, KC managed the research process, undertook some analysis and contributed to the writing, AMc undertook the qualitative evaluation and contributed to the findings, JK contributed to the design and analysis, SS and RT proposed the project and contributed to design, writing and interpretation.

**Competing interests:** Sarah Scobie and Richard Thomson were responsible for commissioning this work on behalf of the National Patient Safety Agency.

**Funding sources:** Funding for the study was provided by the National Patient Safety Agency to the University of Sheffield.

**Copyright statement:** The Corresponding Author has the right to grant on behalf of all authors and does grant on behalf of all authors, an exclusive licence (or non exclusive for government employees) on a worldwide basis to the BMJ Publishing Group Ltd, and its Licensees to permit this article (if accepted) to be published in BMJ editions and any other BMJPGL products and to exploit all subsidiary rights, as set out in our licence (bmj.com/advice/copyright.shtml).

**References**

1. Kohn LT, Corrigan JM, Donaldson MS (Institute of Medicine). To err is human: building a safer health system. *Washington, DC: National Academy Press*, 2000.
2. Department of Health. An organisation with a memory, 2000. [www.dh.gov.uk/assetRoot/04/08/89/48/04088948.pdf](http://www.dh.gov.uk/assetRoot/04/08/89/48/04088948.pdf) (accessed 20 November 2006).
3. Williams SK, Osborn SS. The development of the National Reporting and Learning System in England and Wales, 2001-2005. *Med J Aust* 2006;**184**:S65-S68.
4. Spigelman AD, Swan J. Review of the Australian incident monitoring system. *ANZ J Surg* 2005;**75**:657-61.
5. Agency for Healthcare Research and Quality. Patient Safety Reporting Systems Factsheet, 2001. [www.ahrq.gov/qual/taskforce/hhsrepor.htm](http://www.ahrq.gov/qual/taskforce/hhsrepor.htm) (accessed 20 November 2006).
6. National Patient Safety Agency: National Reporting and Learning System. [www.npsa.nhs.uk/health/reporting/background](http://www.npsa.nhs.uk/health/reporting/background) (accessed 20 November 2006).
7. National Patient Safety Agency. Quarterly National Reporting and Learning System Data Summary, Autumn 2006.

[www.npsa.nhs.uk/site/media/documents/1953\_NRLS\_Data.pdf](http://www.npsa.nhs.uk/site/media/documents/1953_NRLS_Data.pdf) (accessed 20 November 2006).

1. World Health Organization. World Alliance for Patient Safety: WHO draft guidelines for adverse event reporting and learning systems, 2005.

[www.who.int/patientsafety/events/05/Reporting\_Guidelines.pdf](http://www.who.int/patientsafety/events/05/Reporting_Guidelines.pdf) (accessed 20 November 2006).

1. Giles S, Fletcher M, Baker M, Thomson R. Incident reporting and analysis. In: Walshe K, Boaden R, eds. *Patient Safety: Research into Practice. Open University Press*, 2005.
2. Evans SM, Berry JG, Smith BJ, Esterman A, Selim P, O’Shaughnessy J, et al. Attitudes and barriers to incident reporting: a collaborative hospital study. *Qual Saf Health Care* 2006;**15**:39-43.
3. Kingston MJ, Evans SM, Smith BJ, Berry JG. Attitudes of doctors and nurses towards incident reporting: a qualitative analysis. *Med J Aust* 2004;**181**:36-39.
4. Schectman JM, Plews-Ogan ML. Physician perception of hospital safety and barriers to incident reporting. *Jt Comm J Qual Patient Saf* 2006;**32**:337-43.
5. Waring JJ. Beyond blame: cultural barriers to medical incident reporting. *Soc Sci Med* 2005;**60**:1927-35.
6. Johnson CW. How will we get the data and what will we do with it then? Issues in the reporting of adverse healthcare events. *Qual Saf Health Care* 2003;**12 Suppl 2**:ii64-7.
7. World Health Organization. World Alliance for Patient Safety: Project to Develop an International Patient Safety Event Classification. [www.who-ipsec.org](http://www.who-ipsec.org) (accessed 20 November 2006).
8. Department of Health and Health Protection Agency. MRSA surveillance system. [www.hpa.org.uk/infections/topics\_az/hai/mandatory\_report\_2006.htm](http://www.hpa.org.uk/infections/topics_az/hai/mandatory_report_2006.htm) (accessed 20 November 2006).
9. Medicines and Healthcare Products Regulatory Agency: Medical device reporting system. [www.mhra.gov.uk/home/idcplg?IdcService=SS\_GET\_PAGE&nodeId=355](http://www.mhra.gov.uk/home/idcplg?IdcService=SS_GET_PAGE&nodeId=355) (accessed 20 November 2006).
10. Healthcare Commission NHS Staff Survey.

[www.healthcarecommission.org.uk/nationalfindings/surveys.cfm](http://www.healthcarecommission.org.uk/nationalfindings/surveys.cfm) (accessed 20 November 2006).

1. NHS Litigation Authority Risk Management Standards.

[www.nhsla.com/RiskManagement](http://www.nhsla.com/RiskManagement/) (accessed 20 November 2006).

1. Healthcare Commission Hospital Clusters.

<http://ratings2004.healthcarecommission.org.uk/cluster.asp> (accessed 20 November 2006).

1. Hospital Episode Statistics Online: Data on hospital providers.

[www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=212](http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=212) (accessed 20 November 2006).

1. Dr Foster. The Hospital Guide: How Good is My Hospital, 2005. [www.drfoster.co.uk/hospitalreport/pdfs/howGood.pdf](http://www.drfoster.co.uk/hospitalreport/pdfs/howGood.pdf) (accessed 20 November 2006).
2. Agency for Healthcare Research and Quality: Quality Indicators. [www.qualityindicators.ahrq.gov/psi\_overview.htm](http://www.qualityindicators.ahrq.gov/psi_overview.htm) (accessed 20 November 2006).
3. Chatfield C. The analysis of time series: An introduction. *London: Chapman and Hall/CRC*, 2003.
4. StataCorp. Stata Statistical Software: Release 9.0. College Station, Texas: Stata Corporation2006.
5. National Patient Safety Agency. Building a memory: preventing harm, reducing risks and improving patient safety: The first report of the National Reporting and Learning System and the Patient Safety Observatory, 2005. [www.npsa.nhs.uk/site/media/documents/1246\_PSO\_Report\_FINAL.pdf](http://www.npsa.nhs.uk/site/media/documents/1246_PSO_Report_FINAL.pdf) (accessed 20 November 2006).
6. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;**363**:1147-54.
7. Sexton JB, Helmreich RL, Neilands TB, Rowan K, Vella K, Boyden J, et al. The Safety Attitudes Questionnaire: psychometric properties, benchmarking data, and emerging research. *BMC Health Serv Res* 2006;**6**:44.

**10.3 Publication 3**

Dean JE, **Hutchinson A**, Hamilton Escoto K, Lawson R. Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway. *BMC Health Services Research* 2007;7:89

Doi:10.1186/1472-6963-7-89

This article is available from [www.biomedcentral.com/1472-6963/7/89](http://www.biomedcentral.com/1472-6963/7/89)

**Using a multi-method, user centred, prospective hazard analysis to assess care quality and patient safety in a care pathway**

Joanne E Dean1\*§, Allen Hutchinson1\*, K Hamilton Escoto2, R Lawson3

1Section of Public Health, ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, UK, S1 4DA

2 Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA

3 Department of Respiratory Medicine, University of Sheffield, Sheffield, UK

\*These authors contributed equally to this work

§Corresponding author

Email addresses:

JED: [j.e.dean@sheffield.ac.uk](mailto:jane@darwin.co.uk)

AH: [Allen.Hutchinson@sheffield.ac.uk](mailto:Allen.Hutchinson@sheffield.ac.uk)

KHE: [kjhamilton@wisc.edu](mailto:kjhamilton@wisc.edu)

RL: [Rod.Lawson@sth.nhs.uk](mailto:Rod.Lawson@sth.nhs.uk)

**Abstract**

**Background**

Care pathways can be complex, often involving multiple care providers and as such are recognised as containing multiple opportunities for error. Prospective hazard analysis methods may be useful for evaluating care provided across primary and secondary care pathway boundaries. These methods take into account the views of users (staff and patients) when determining where potential hazards may lie. The aim of this study is to evaluate the feasibility of prospective hazard analysis methods when assessing quality and safety in care pathways that lie across primary and secondary care boundaries.

**Methods**

Development of a process map of the care pathway for patients entering into a Chronic Obstructive Pulmonary Disease (COPD) supported discharge programme. Triangulation of information from: care process mapping, semi-structured interviews with COPD patients, semi-structured interviews with COPD staff, two round modified Delphi study and review of prioritised quality and safety challenges by health care staff.

## **Results**

Interview themes emerged under the headings of quality of care and patient safety. Quality and safety concerns were mostly raised in relation to communication, for example, communication with other hospital teams. The three highest ranked safety concerns from the modified Delphi review were: difficulties in accessing hospital records, information transfer to primary care and failure to communicate medication changes to primary care.

**Conclusions**

This study has demonstrated the feasibility of using mixed methods to review the quality and safety of care in a care pathway. By using multiple research methods it was possible to get a clear picture of service quality variations and also to demonstrate which points in the care pathway had real potential for patient safety incidents or system failures to occur. By using these methods to analyse one condition specific care pathway it was possible to uncover a number of hospital level problems. A number of safety challenges were systems related; these were therefore difficult to improve at care team level. Study results were used by National Health Service (NHS) stakeholders to implement solutions to problems identified in the review.

**Background**

There is a substantial literature from engineering and safety critical industries on the use and methods of prospective hazard analysis. [1] A more limited literature relates to health care [2] where adapted methods such as Health Care Failure Modes and Effects Analysis [3] have been used to assess risks in high risk systems such as blood transfusion. [4] Other methods, such as socio-technical probabilistic risk analysis [5], are being used in aspects of medication safety or have been used in the study of complex environments such as anaesthesia. [6]

Where they have been used in health care, most of these risk analyses have taken place in hospital-based settings, focussing on well-defined care processes. Although retrospective methods such as Root Cause Analysis are frequently used in healthcare, it is less usual to find accounts of the use of prospective hazard analysis methods across boundaries between primary and secondary care or that take account of the views of patients in determining where hazards may lie. Yet care pathways that cross the boundary between community and hospital care are recognised to contain opportunities for error, both latent and active. [7]

Hence the primary purpose of this study was to investigate the feasibility and value of using mixed research methods to inform a prospective hazard analysis of risks in a care pathway that crosses primary and secondary care boundaries.

The Sheffield supported discharge programme for people with COPD was used as the study setting. The programme’s aim was to reduce hospital length of stay by providing specialist, hospital-based, nursing care at home until the acute episode resolved.

Evidence suggests that early supported discharge may be a safe (using outcomes of readmission rates and mortality) and perhaps cost effective means of improving care for some people with acute exacerbations of chronic obstructive pulmonary disease. [8][9][10][11] Moreover, patients and their carers seem to prefer the option of early homecare where possible, [12] and its use is recommended in the UK National Institute for Clinical Excellence guidelines on chronic obstructive pulmonary disease management. [13]

Safety hazards in care pathways are likely to be quantified by a set of measures broader and more responsive than readmission and mortality rates. [11] For example, Reason [7] has demonstrated how safety is a dynamic process where risks change over time and where some risks are ‘latent’ (and difficult to foresee) while others are active and may be everyday occurrences. Dekker [14] graphically explores how departures from the routine practice envisaged in a pathway eventually become the routine. Cook [15] characterises the risks of everyday clinical practice as working within a ‘safety envelope’, where production pressures (such as the need to reduce length of stay to release bed space) can push clinical practice through the margin of safety into the unpredictable territory where accidents happen.

Recent reviews of quality evaluation methods [16][17][18] have identified a range of possible approaches for assessing quality and safety in care pathways, including methods for seeking users views and the use of Delphi methods that might be used to gain professional views on risks and potential solutions [19]. Woods and Cook [20] have drawn on the application of safety

science in anaesthesia to develop a proactive check list that can be used to seek out points were safety is more vulnerable (Figure 1).

Pursuit of second stories [20] is a pre-requisite of understanding how the care pathway is actually structured, since the initial plan for any care pathway is rarely followed in every detail when put into practice. Clarity about the focus of a hazard analysis can also be gained by using standardised work process methods to establish a process map of the care pathway. [21][22]

This paper reports the results of a prospective hazard analysis of care quality and patient safety in a COPD supported discharge care pathway.

**Methods**

#### Mapping the care process

Mapping of the care pathway, from admission to hospital to the point of discharge from the supported discharge programme, was undertaken iteratively through 8 one-to-one interviews and three meetings with hospital and community (Primary Care Trust) staff who were involved in designing, implementing and providing the Sheffield COPD supported discharge programme. A single observer accompanied hospital-based nursing staff during three domiciliary visits to seek further information about the supported discharge process. Finally, a joint meeting was held with hospital staff to agree the contents of the care pathway map. A standardised format [21][22] was used to create the process map. The key care decision points and processes identified from the map were subsequently the focus for the interviews with patients and staff about the process, outcome and safety of care.

Additionally, the supported discharge programme record keeping process was reviewed to determine how the information from the supported discharge scheme was retained in the hospital paper based records system.

#### Review of quality and safety of care

Interviews

The interview schedules were based on a standardised work systems framework drawn from process engineering [23] and comprised six main themes: process understanding, work system understanding, communication, documentation, problems and suggestions for improvements and quality of care received (patient’s only).

Semi-structured interviews were undertaken with five hospital nursing staff and two medical hospital staff involved with providing the supported discharge service.

Sixteen patients were selected for interview by nursing staff, through purposive sampling. All patients had recently been admitted to hospital with an exacerbation of symptoms of chronic obstructive pulmonary disease, had been in the 14 day supported discharge programme and had subsequently been discharged to primary care. Semi structured interviews were undertaken with all 16 patients.

**Analysis of the interview data**

The analysis of the interview data was undertaken in 3 stages.

*Stage 1*

Interview data from staff and patients were analysed using FRAMEWORK, [24] an explicit, structured method of qualitative data analysis employing 5 distinct but interconnected stages in a systematic process. These stages are familiarisation; identifying a thematic framework; indexing; charting; mapping and interpretation.

For each stage in the analysis, one analyst (JD) developed a first draft of the results. The process was reviewed by a second analyst (AH) and the findings either confirmed or modified through joint discussion.

*Stage 2*

Information from the thematic FRAMEWORK analysis was plotted on to the process of care map, together with information from the documentary review of record keeping, to identify points in the care process where quality and safety might be variable. The findings were reviewed by a second analyst (AH) and these were then agreed or modified in joint discussion.

*Stage 3*

The 23 sub-themes arising from the first two analysis phases were brought together to look for similarities and differences. A final set of six main themes were developed, together with a number of specific safety vulnerabilities.

**Assessing relative safety risk**

Analysis of the interview data from patients and staff highlighted specific areas or care pathway points where it was felt that care quality and or patient safety could potentially be compromised. A modified, two–stage, questionnaire based Delphi approach was used to obtain staff views about what worked well and what worked less well in eight of the identified pathway areas. These were:

Re-admission management; clinical organisation; communication within the chronic obstructive pulmonary disease team; patient knowledge about the supported discharge programme; communication with the hospital bed bureau; information priorities; communication with primary care.

In the first round, 7 supported discharge programme staff and 9 primary care staff who were members of the local Joint Care Planning Group (including: 2 General Practitioners, 2 Primary Care Chronic Disease Nurses, 1 Respiratory Nurse Specialist and COPD Primary Care Manager, 1 PCT Service Development Manager, 1 PCT Health Improvement Manager for Chronic Disease, 1 PCT Director of Public Health, 1 Occupational Therapist Service Manager) and the Manager of the Medical Directorate of the local Health Care Trust were contacted.

In a second round, all of the data sets were used to identify specific potential safety vulnerabilities within the supported discharge programme. Respondents were asked to rank each of these safety vulnerabilities in terms of risk to patients, using a visual analogue scale (0 = not a safety problem, 10 = significant safety challenge). Additionally, respondents were asked for their ideas on how each of the safety risks could be resolved or their impact reduced.

Finally, members from the chronic obstructive pulmonary disease team and the local Joint Care Planning Group (responsible for commissioning the chronic obstructive pulmonary disease programme from the hospital) used a Failure Mode Effect Analysis [3] approach to discuss possible solutions for three of the most highly ranked safety problems,

**Research governance**

The study was reviewed for research governance purposes by the Sheffield Health and Social Care Research Consortium and Sheffield Teaching Hospitals Trust. Ethics review was undertaken by the North Sheffield Local Research Ethics Committee.

**Results**

#### Mapping the care process

The final version of the pathway reflected a complex care process that was different between the two hospital sites within the health care Trust. Although the quality and safety of care results focus mainly on the detail of the structured discharge process, this could only be understood in the context of the whole care pathway, including admission routes and transfers between units within the hospital sites (see Figure 2).

**Results from qualitative analysis of interview data**

**Quality of care:** Variation in organisation and clinical practice across hospital sites

There was provision of a differential service between the two hospital sites that provided the supported discharge service. Some COPD supported discharge services provided by one site were not provided by the other. For example, one site had a service agreement with social services to provide a home care review within 48 hours of a request. The other did not. This variation resulted delayed admission to the supported discharge programme for one site.

**Quality of care:** Patient satisfaction

Patients preferred to be cared for at home rather than at hospital (14/16) and they felt more confident in returning home with the knowledge that a nurse would visit them on a daily basis. Communication between staff and patients was rated highly by most patients (14/16). One patient and family was upset over the way they perceived the supported care process had been handled in a recent short-term admission to hospital.

**Patient safety:** *Communication with other hospital teams*

Under the supported discharge scheme, re-admission to hospital could be initiated by the respiratory nurse specialists or by the patient by telephoning the hospital bed bureau. The purpose of the bed bureau was to manage the flow of hospital admissions, ensuring that patients are admitted to the relevant hospital site. This method of re-admission was in addition to traditional re-admission routes via telephoning the emergency ambulance service or the patient’s General Practitioner. There was confusion about this re-admission process on various levels, particularly on the part of hospital staff who were not part of the supported discharge programme, for example in the acute admissions department, which sometimes resulted in delays in admission. On other occasions, failure by hospital teams to inform the COPD staff that a re-admission had taken place resulted in patients not being seen as quickly as intended and, in turn, extended the patient’s time in a hospital bed. Patients had experience of failing to be re-admitted, having exercised their right to ask for re-admission via the hospital bed bureau. Because bed bureau telephone lines were often engaged, nursing staff sometimes had difficulty in communicating with the bureau when attempting to arrange emergency admission from the community.

*Patient safety: Communication between the chronic obstructive pulmonary disease team and primary care*

For a variety of reasons, up to date information about discharged patients sometimes seemed not to reach relevant primary care staff. Nursing staff rarely managed to speak to primary care staff. The COPD staff had to make contact via the practice receptionist using the same telephone line as patients. This telephone line was often busy and on most occasions the COPD staff could only leave a message as the health professional they wished to speak to was busy. Instead, the COPD staff regularly faxed and posted documentation to the practice. However, this did not always reach the relevant staff member. Some difficulties were structural. For instance, faxed copies of discharge summaries failed to reach primary care because, in some city general practices, fax machines were switched off on afternoons when the practices are closed.

*Patient safety: Telephone communications*

Technical difficulties in telephone communications occurred regularly between staff and between staff and patients. The supported discharge programme staff spent over half of their time in the community and patients contact staff via mobile phone. But hospital rules require that mobile phones are switched off when staff are on hospital premises so the usual route for patient to staff contact was not always available.

One of the criteria for selection of patients for the supported discharge programme was that there should be an outgoing call landline in the patient’s house so that a patient could call for assistance if their condition deteriorated. But in practice the community staff came across a variety of telephone difficulties, including land lines with no outgoing call facility, no landline and mobile phones which were uncharged or without credit.

*Patient safety: Access to medical records*

A complex flow process for the paper-based patient records in the post-discharge period resulted in limited access to patient records when patients attend consultant/physician led hospital outpatient clinics. This was particularly a problem for one of the sites because the hospital records were stored on the other hospital site and administrative staff found it difficult to source the records for the clinic. This led to records routinely not being sourced for the clinic because of the pressures it placed on staff time. Therefore it was not unusual for outpatient appointments to be held without patient records of the prior admission.

**Results from the Delphi study**

Twelve potential safety challenges were identified from the interviews and confirmed in the first round, with no additions. In the second round these were ranked by eight NHS staff. Descriptors and their relative rankings are displayed in Table 1. Respondents offered their ideas on solutions for each of the 12 safety challenges.

**Failure Mode and Effect Analysis**

Five senior NHS staff attended the final Failure Mode and Effect Analysis meeting to discuss the results, choosing to seek solutions to safety problems 1, 2 and 3 combined (because they both related to primary care teams) and 5 (Table 1). Participants did not review service problems during periods of sickness absence (problem 4) since this difficulty was thought to have been resolved.

However, pressure of time commitments on the NHS staff meant that they were not able to commit to the effort required for a formal Failure Mode and Effect Analysis and all that could be achieved was a structured discussion of the safety challenges and recording of proposed solutions.

**Discussion**

The study has demonstrated the feasibility of using mixed methods to review the quality and safety of care provided through a care pathway. For much of the time the reviewed care pathway worked reasonably well and safely and on the whole the 16 interviewed patients were satisfied with the supported discharge process and the care they received.

Nevertheless the results also highlighted aspects of health care organisation where patients were potentially vulnerable to poor quality or unsafe care. Most health care systems contain latent safety risks, in which embedded organisational factors and local workplace factors can conspire to breach the defences of the system. [7] [25]. Since it is still uncommon for care pathways to be re-designed with specific attention to patient safety, including carrying out some form of prospective hazard analysis, it is not surprising that a number of safety vulnerabilities were found.

Some vulnerabilities were organisational and were exacerbated by the changing dynamics of the health care system, such as the recent merger of two hospitals. But such system weaknesses require a range of organisational developments to improve the service, many of which were outside the direct area of influence of the COPD team. These included achieving agreement with admission teams and the bed bureau over how the structured programme facilities for emergency admission might work. A particular problem seemed to be a failure to keep informed those who were still in medical staff in training positions (and who had short term appointments) about the admission processes of the supported discharge programme.

The challenge of accessing paper based medical records for some of the follow-up clinics, and the eventual resigned acceptance of the status quo after failed attempts to rectify the problem, is an example of what Dekker [14] refers to as ‘deviations from the routine becoming the routine’. Electronic records may be the solution here but, in the short term, local workplace initiatives are required to improve the safety of care provided. These changes did not seem to be within the influence of the chronic obstructive pulmonary disease team – this is a hospital level problem.

Even telephone communications between staff and patients and patients and the hospital had unforeseen problems that were difficult to deal with. Having an outgoing call landline was a criterion for selection for the supported discharge programme – some patients evidently slip through this selection criterion. Despite the fact that the communication problems were well known to the nursing staff, nothing had been done to tighten the patient selection criteria for the supported discharge programme.

There are methodological limitations to this study. It was undertaken in two sites (within one hospital Trust) so the findings may not be generalisable. However, the feasibility of methods may well be, even though it was sometimes difficult to keep all parties engaged in the research throughout the study period. It had been intended to undertake a prospective Health Care Failure Mode and Effect Analysis [3] as the final stage but this was not possible as a number of staff were unable to attend the meeting. The nature of health care meant that some contributors had to cancel on the day of the meeting.

**Conclusions**

By triangulatinginformation from a detailed mapping of the care pathway, from views and from concerns of users, and by ranking problems by potential severity of impact on care quality and patient safety, it was possible to get a clear picture of service quality variations in the supported discharge programme. It was also possible to demonstrate which points in the care pathway had real potential for patient safety incidents or system failures to occur. Most of the variations in quality of care and care organisation resulted from system deficits and some required coordinated effort to achieve improvements.

Safety can be defined in more ways than mortality and admission rates and this study demonstrates how much of the potential for safety incidents at the individual level is embedded in the design and the actuality of the care pathway and its processes. Taking into account the reality of finding time in the lives of busy healthcare professionals, whose first response is to immediate healthcare pressures, the use of other methods of gathering information about prospective hazards, such as interviews of patients and staff and the use of Delphi methods to capture additional data, may be a more successful alternative to a full-scale, formal prospective hazard analysis.

**Competing interests**

RL is one of the clinicians providing the supported discharge service.

**Authors' contributions**

AH conceived of the study, obtained funding, oversaw the running of the study, analysed data and contributed to the development of the care pathway and the writing of the manuscript. JED obtained ethics approval, developed data collection materials conducted fieldwork, analysed data and contributed to the development of the care pathway and the writing of the manuscript. KHE contributed to the literature searches, development of the care pathway, development of interview schedules and the writing of the manuscript. RL was involved in the conception of the project, project meetings, data collection and contributed to the development of the care pathway and the writing of the manuscript.

**Acknowledgements**

We would like to acknowledge all staff involved in the delivery and organisation of care for patients in the Sheffield COPD supported discharge programme.

Partial funding was provided by Sheffield Health and Social Care Research Consortium

**References**

1. Kirwan B. Human error identification in human reliability assessment. Part 2: Detailed comparison of techniques. Applied ergonomics 1992;23; 371-381

2. McDonagh JE, Solomon R, Petosa L. Qualitative improvement and prospective hazard analysis models: deciphering a new Tower of Babel in *Patient Safety,* The National Academies Press, 2004

3. DeRosier J, Stalhandske E, Bagian J, Nudell T. Using health care failure mode effect analysis. *J Qual Improvement* 2002 ;28: 248-267

4. Burgmeier J. Failure mode and effect analysis: an application in reducing risk in blood transfusion. Joint Commission Journal on Quality Improvement. 28(6): 331-9, 2002

5. Marx DA. Slonim AD. Assessing patient safety risk before the injury occurs: an introduction to sociotechnical probabilistic risk modelling in health care. Qual Saf Health Care. 12 Suppl 2: ii33-8, 2003

6. Pate-Cornell E, Murphy D, Lakats L, Gaba D. Patient risk in anaesthesia: probabilistic risk analysis and management improvements. Ann Operations Research 1996;67; 211-233

7. Reason J. *Managing the risks of organisational accidents*. Aldershot: Ashgate 1997

8. Davies L, Wilkinson M, Bonner S *et al*. ‘Hospital at home’ versus hospital care in patients with exacerbations of chronic obstructive pulmonary disease: prospective randomised controlled trial. *BMJ* 2000;321: 1265-1268

9. Skwarska E, Cohen G, Skwarski KM *et al*. Randomised controlled trial of supported discharge in patients with exacerbations of chronic obstructive pulmonary disease. *Thorax* 2000;55: 907-12.

10. Cotton MM, Bucknall CE, Dagg KD *et al*. Early discharge for patients with exacerbations of chronic obstructive pulmonary disease: a randomised controlled trial. *Thorax* 2000;55: 902-6

11. Ram FRS, Wedzicha J, Wright J *et al*. Hospital at home for patients with chronic obstructive pulmonary disease: a systematic review. *BMJ* 2004; 329:315-8

12. Ojoo JC, Moon T, McGlone S *et al*. Randomised controlled trial of home based care of patients with chronic obstructive pulmonary disease. *Thorax* 2002;57: 167-9

13. *Thorax* 2004;59(Suppl 1):i131-i156

14. Dekker S. *The field guide to human error investigations*. Aldershot: Ashgate 2002

15. Cook RI. A brief look at gaps in the continuity of care and how practitioners compensate for them.

<http://www.ctlab.org/pdf%20files/brieflookatgapsBackpageRevf.pdf>

Accessed 5/5/05

16. Ovretveit J, Gustafson D. Using research to inform quality programmes. *BMJ* 2003;**326**: 759-761

17. Campbell SM, Brasspenning J, Hutchinson A *et al*. Research methods used in developing and applying quality indicators in primary care. *BMJ* 2003;326: 816-819

18. Wensing M, Elwyn G. Methods for incorporating patients’ views in health care. *BMJ* 2003;326: 877-879

19. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000;32: 1008-1015

20. Woods DD, Cook RI. Nine steps to move forward from error. *Cognition Technology &Work* 2002;4:137-144

21. Lighter DE and Fair DC. *Principles and methods of quality management in health care*. Gaithersburg: Aspen 2000

22. NHS Modernisation Agency. Improvement Leaders guide to process mapping, analysis and design. <http://www.modern.nhs.uk/improvementguides/process/> accessed 6/12/04

23. Smith MJ, Sainfort PC. Balance theory of job design for stress reduction. *Int J Ind Ergon* 1989;4: 67-79

24. Ritche J, Spencer L. *Qualitative data analysis for applied policy research* in Bryman A, Burgess RG (eds) *Analysing Qualitative Data,* London: Routledge, 1994 Pp173-194.

25. Rasmussen J. Human errors: a taxonomy for describing human malfunction in industrial installations. *J Occupational Accidents* 1982;4: 311-333

# **Tables**

**Table 1- Type of safety problem ranked by mean score on a visual analogue scale**

|  |  |  |
| --- | --- | --- |
| **N** | **Safety problem descriptor** | **Mean score/ 10** |
| 1 | Routine difficulties with access to medical records in post discharge clinics leads to decisions being made without adequate background information | 6.9 |
| 2 | For a variety of reasons, information about discharged patients sometimes does not reach relevant primary care staff | 6.8 |
| 3 | Patients are at risk when medication changes during admission are not communicated to primary care | 6.0 |
| 4 | The service is vulnerable during periods of staff sickness, which may also affect staff morale | 5.5 |
| 5 | Difficulty in communicating with the bed bureau can put patients at risk | 5.4 |
| 6 | The provision of a differential service across the two hospitals may lead to a variation in the quality of the care provided | 5.2 |
| 7 | Some primary care staff appear to be unsure of the aim of the supported discharge programme, and of the care provided | 5.0 |
| 8 | Patients are at risk when patients do not bring their home care treatment/record with them on re-admission | 5.0 |
| 9 | Making and keeping hospital appointments can be a problem | 4.9 |
| 10 | Lack of clarity on the part of non-COPD Hospital Staff about the re-admission process leads to quality variation and admission delays, misdirection of patients and inefficiencies | 4.6 |
| 11 | Technical difficulties with telephone communications between staff and between staff and patients is a possible safety risk | 4.5 |
| 12 | Quality variation and inefficiencies occur because the COPD Supported Discharge Programme does not have a high priority, compared with other hospital services | 3.8 |

**Figures**

## Figure 1 - Adapted from Woods DD and Cook RI, Nine steps to move forward from error [20]

## 

1. Pursue second stories beneath the surface to discover multiple contributors to a safety hazard
2. Understand work as performed at the sharp end of the system (and not depend on being told how it is intended to be)
3. Search for systemic safety vulnerabilities
4. Study how practice creates safety

5. Examine how change to the care pathway will produce new vulnerabilities and paths to failure. legend text

.

# 

**10.4 Publication 4**

**Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Rantell K, Campbell MJ, Nicholl J, Irwin P. Assessing quality of care from hospital case notes: comparison of reliability of two methods. Quality and Safety in Health Care 2010;19:e2.

doi:10.1136/qshc.2007.023911

**Please note:** Because it has not been possible to contact one of the authors of this study in order to seek their permission to publish, this publication can only be accessed through electronic search systems. The doi address may help in this regard.

**10.5 Publication 5**

**Hutchinson A**, Coster JE, Cooper KL, McIntosh A, Walters SJ, Bath PA, Pearson M, Young TA, Rantell K, Campbell MJ, Ratcliffe J. Comparison of case note review methods for evaluating quality and safety in health care. Health Technology Assessment 2010;14(10):1-170

doi:10.3310/hta14100

also at: <http://eprints.whiterose.ac.uk/10704/>

**Comparison of case note review methods for evaluating quality and safety in health care**

**Authors:**  Allen Hutchinson, Joanne Coster, Katy Cooper, Aileen McIntosh, Stephen Walters, Peter Bath, Michael Pearson, Tracey Young, Khadija Rantell, Michael Campbell, Julie Ratcliffe

**Section of Public Health, ScHARR, University of Sheffield**

A Hutchinson, J E Coster\*, K L Cooper, A McIntosh (\* Formerly known as Dean)

**Section of Health Services Research, ScHARR, University of Sheffield**

S J Walters, K Rantell, M J Campbell,

**Section of Health Economics and Decision Sciences, ScHARR, University of Sheffield**

T A Young, J Ratcliffe

**Department of Information Studies, University of Sheffield**

P A Bath

**Clinical Effectiveness and Evaluation Unit, Royal College of Physicians**

M Pearson (now **Department of Medicine, University of Liverpool**)

Corresponding Author

Professor Allen Hutchinson

Section of Public Health

ScHARR, University of Sheffield

Regent Court, 30, Regent St

Sheffield, S1 4DA

Email: [Allen.Hutchinson@sheffield.ac.uk](mailto:Allen.Hutchinson@sheffield.ac.uk)

Tel; +44 (0)1142220811 **Contents**

Preface: aims, general approach to the study and structure of the report…………. 4

Acknowledgements ……………………………………………………………………….…6

**Summary**……………………………………………………………………………………..7

**Assessing quality of care from hospital case notes: comparison of reliability and utility of holistic (implicit) and criterion based (explicit) methods**

Background………………………………………………………………………………….10

Study aim and research questions ……………………………………………...13

Methods ……………………………………………………………………………………..14

Choice of conditions, review methods, settings and staff …………………….14

#### Data capture methods………………….........………………………………...19

Analysis methods …………………………………………………………………24

Research ethics review and governance………………………….……………30

Results ……………………………………………………………………………………...31

Analysis of holistic review scale scores ………………………………………...32

Analysis of review criterion based scores ……………………………………...36

Comparison of holistic and criterion based methods ………………............37

Thematic analysis of holistic textual data ………………………………………41

Resource implications….…………………………………………………………49

Quality of care – hospital mortality group relationship ………………………..50

Feedback to study hospitals and teams ………………………………………..57

Conclusions .………………………………………………………………………………..58

Implications for reviewing quality of care ...…………………………………….63

**What is the relationship between information on quality of care from case notes and hospital level outcomes of care?**

Background ………………………………………………………………………………...65

Methods …………………………………………………………………………………….67

Choice of settings, review methods and staff …………………………............67

Selection of outcome data ...……………………………………………………..72

Analytic approach …………………………………………………………………73

Results …………….………………………………………………………………………..74

Relationship between outcome variables and higher mortality and lower mortality groups of hospitals……………………………………………………...74

Correlations between holistic scale scores and hospital level outcome variables ……………………………………………………………………………78

Correlations between criterion based scores and hospital level outcome variables..………………………………………………………………………......85

Outliers in the analysis …………………………………………………………...87

Conclusions .………………………………………………………………………………..90

Overview of findings ...……………………………………………………………93

**Future research agenda**………………………………………………………….......................................95

**References** ..……………………………………………………………………………….96

**Appendices**

Appendix 1: COPD review criteria ……………………………………………………...101

Appendix 2: Heart failure review criteria ……………………………………………….111

Appendix 3: Validity of review criterion questionnaire (COPD) ……………………..119

Appendix 4: Screen shot of a holistic review data collection page …………………126

Appendix 5: Reviewer training scenarios to assist in recognising variation in care quality using holistic review ……………………………………………………………..127

Appendix 6: Example of a reviewer feedback report ……………………………......130

Appendix 7: COPD – correlations between holistic mean overall scale scores and outcome variables ……………………………………………………………………….153

Appendix 8: Heart Failure – correlations between holistic mean overall scale scores and outcome variables …………………………………………………………………..154

Appendix 9: COPD – correlations between holistic mean phase scale scores and outcome variables ………………………………………………………………………..155

Appendix 10: Heart Failure – correlations between holistic mean overall scale scores and outcome variables …………………………………………………………………..156

Appendix 11: COPD – correlations between holistic mean criterion scores and outcome variables ………………………………………………………………………..157

Appendix 12: Heart Failure – correlations between mean criterion scores and outcome variables ………………………………………………………………………..158

Appendix 13: Comparison of holistic and criterion based review methods using structured clinical records in stroke care……………………………………………….159

Appendix 14: The place of trigger tool methodology in case note review for quality and safety………………………………………………………………………………….173

**Preface: aims, general approach to the study and structure of the report**

The study had two main aims which were agreed in response to a call for proposals from the Department of Health Methodology Research Programme. These aims were:

1. To compare the validity and reliability of two methods of case note review of quality and safety of care. That is, to explore which of two methods of case note review, holistic (implicit) review or criterion based (explicit) review, is the most effective method of reviewing quality of care, and under what circumstances and by which type of staff. Methodological questions include a comparison of reliability of holistic (implicit) and criterion based (explicit) methods.

2. To investigate whether there appears to be a link between the quality of medical care, as recorded in case notes, and the outcome of hospital care, for two chronic medical conditions. Methodological questions include an exploration of how implicit and explicit case note review might be used to explore the relationship between process of care and risk adjusted outcomes.

Since the aims are linked but address two different aspects of case note review, the methodology of case note review and process/outcome relationship, this report is presented in two parts, while the overall summary is presented as a whole. Figure 1 shows where the linkages lie between the two studies, particularly through the choice of review methods, type of reviewers and methods of selection of hospitals.

Additionally, two small studies were also commissioned as part of the programme of work. One uses the case note review methods to explore their value in the context of structured record keeping, using stroke care as an example. The second study explores the literature relating to the use of trigger tools when reviewing paper-based case notes for quality and safety. These studies are presented in Appendices 13 and 14 respectively.



**Acknowledgements**

This study was undertaken by two partner organisations, the University of Sheffield - the School of Health and Related Research (ScHARR) and the Department of Information Studies - and the Royal College of Physicians Clinical Evaluation and Effectiveness Unit (CEEu). An important feature of the study was the contact made with clinical teams in over 30 hospitals in England, many of whom went on to contribute to the study. CEEu staff took a major part in enabling the Sheffield team to contact the hospitals and their specialist teams, as well as providing expertise in case note review and quality of care methodology.

Karen Beck provided administrative support at the University of Sheffield throughout the project and her contribution was exceptionally helpful. In particular her substantial contribution to the development of data collection software was invaluable.

Jon Nicholl provided valuable methodological advice in the early part of the study.

We especially wish to acknowledge the enthusiasm and assistance of reviewers and all their colleagues in the study hospitals, without whom this study could not have taken place.

**Comparison of case note review methods for evaluating quality and safety in health care**

**Summary**

**Purpose**

The purpose of the first part of the study was two-fold. First, to determine which of two methods of case note review provide the most useful and reliable information for reviewing quality and safety of care, and for what purpose. Second, to determine the level of agreement within, and between, groups of healthcare professionals (doctors, nurses and other clinically trained staff, and non-clinical audit staff) when they use the two methods to review the same record.

The results were also expected to influence the methods of data capture for the second part of the study, which explored the process-outcome relationship between holistic and criterion based quality of care measures (process measures) and hospital level outcome indicators, grouped by mortality level.

**Methods**

In the first part of the study, retrospective multiple reviews of 684 case notes were undertaken using both holistic (implicit) and criterion based (explicit) review methods. Quality of care measures included evidence based review criteria and a quality of care rating scale. Textual commentary on the quality of care was provided as a component of holistic review. Data collection was conducted in nine randomly selected acute hospitals in England by hospital staff trained in case note review. These local review teams comprised combinations of three staff types: doctors (n=16), specialist nurses (n=10) and clinically trained audit staff (n=3) (n=13 in total), and non-clinical audit staff (n=9).

During the second part of the study, process (quality and safety) of care data were collected from the case notes of 1565 people with either chronic obstructive pulmonary disease (COPD) or heart failure in 20 randomly selected hospitals in England. Doctors collected criterion based data from case notes and used implicit review methods to derive textual comments on the quality of care provided and score the care overall.

**Analysis methods**

Intra-rater consistency, inter-rater reliability between pairs of staff using Intra-class Correlation Coefficients, completeness of criterion data capture, within and between-staff group comparison and between review method comparison. To explore the process/outcome relationship, a range of publicly available health care indicator data was used as proxy outcomes in a multi-level analysis.

**Results**

A total of 1473 holistic reviews and 1389 criterion based reviews were undertaken in the first part of the study.

When same staff type reviewer pairs/groups reviewed the same record, holistic scale score inter-rater reliability was moderate within each of the three staff groups (Intra-Class Correlation Coefficient [ICC] 0.46 – 0.52) and inter-rater reliability for criterion based scores was moderate to good (ICC 0.61 – 0.88). When different staff type pairs/groups reviewed the same record, agreement between the reviewer pairs/groups was weak to moderate for overall care (ICC 0.24 to 0.43).

Comparison of holistic review score and criterion based score of case notes reviewed by doctors and by non-clinical audit staff showed a reasonable level of agreement between the two methods (p-value for difference 0.406 and 0.223 respectively), although results from all three staff types showed no overall level of agreement (p-value for difference 0.057).

Detailed qualitative analysis of the textual data provided by reviewers indicated that the three staff types tended to provide different forms of commentary on quality of care, although there was some overlap between non-clinical audit staff and the nursing group and between the nursing group and the doctors. Thus the non-clinical audit staff mainly reported facts from the case notes. Nurses and clinical audit staff provided commentaries that were mainly about process of care, together with some implicit judgements about the quality of care provided. Information from the doctors tended to be more focussed on technical aspects of care, making rather more explicit judgements on quality of care.

In the process/outcome study there generally were high criterion based scores for all of the hospitals while there was rather more inter-hospital variation between the holistic review overall scale scores. Rich textual commentary on the quality of care verified the holistic scale scores. While there were trends towards hospitals that had lower mortality also having higher quality of care scores, none of these differences were statistically significant. There was only limited correlation between the outcome indicators and the criterion based or holistic scale scores for either condition across the 20 hospitals.

**Conclusions**

Using a holistic approach to review case notes, groups of the same staff type can achieve reasonable repeatability within their professional groups when asked to rate quality of care on a scale. But there is little agreement between the three staff types when using holistic review methods to rate quality of care for the same clinical record, possibly because the different staff types are exploring different aspects of quality of care, as the qualitative analysis suggests.

All three staff groups have reasonable to high levels of consistency when using criterion based review and, because there tend to be low levels of missing values in the data collected by all three staff types, there is little to chose between the staff groups in terms of reviewer effectiveness.

When the same clinical record was reviewed by the doctors, and by the non-clinical audit staff, using first holistic and then criterion based methods, there is no significant difference between the assessments of quality of care generated by the two methods. This suggests that although the two methods are exploring quality of care differently, they can allow similar levels of quality ratings to be made. When measuring quality of care from case notes, therefore, consideration needs to be given to three important factors:- the method of review, the type of staff to undertake the review and the methods of analysis available to the review team.

It is likely that review of quality of care can be enhanced by using a combination of both criterion based (explicit) methods and structured holistic (implicit) methods, which will identify both evidence based elements of care and the nuances of care that are almost always a component of care in long term conditions. Free textual commentary on the quality of care provided is a valuable asset in judging care, but it is complex to analyse and is likely to remain as a research tool in this field of health care evaluation.

Experience with the review methods shows that variation in quality of care can be identified from a combination of holistic scale scores and textual data review – this combination providing a rich means of understanding the outcome of care on an individual patient basis.

Although there are some correlations between quality of care scores and hospital level outcome data, there is no clear relationship between the process of care and hospital level outcomes for the two indicator conditions in this study. This probably reflects the complexity of the process/outcome relationship at the group level. Available hospital level outcome indicator data are probably insufficiently sensitive to reflect the quality of care recorded in patient case notes. Furthermore, high quality care may be given even when the patient’s outcome is poor, and vice versa. These findings may be pointing to process measures as being more useful than outcome measures when reviewing the care of people who have chronic disease or multiple conditions. **Assessing quality of care from hospital case notes: comparison of reliability and utility of holistic (implicit) and criterion based (explicit) methods**

**Background**

Review of the quality of care as described in written case notes has become a standard means of assessing variation from quality standards and for identifying adverse incidents, either concerning individuals or groups of patients.

Quality of care is currently assessed from clinical records by collecting data using two principal approaches: holistic review (sometimes called implicit review) and criterion based (explicit) review. Both of these approaches have recognised strengths and weaknesses, whether they are being used for performance monitoring and assessment, or in a research setting.

Although attempts to systematise the review of quality of care began nearly a century ago with the work of Codman in 1912,[1] much of the development of case note review methodology began in North America in the 1970s with the work of the Peer Review Organisations, which used implicit review methods (sometimes called ‘holistic methods’) to determine variations in the standards of care provided by hospitals.[2] Subsequently, variants of these holistic (implicit) methods to review the quality of care of hospital patients were used as the basis for determining adverse event and medical error rates in three large epidemiological studies in New York State,[3] Australia[4] and in Colorado and Utah.[5] Holistic review was subsequently widely used in clinical audit in the UK.

Clinical staff in the UK are accustomed to looking through a set of patient records in order to form an opinion on the quality of care delivered. This holistic approach uses professional judgement and has the advantage that it requires no prior assumptions about the individual case, can be applied to any condition, can extend to examining any aspect of care, and, at least in experienced hands, may be relatively quick to perform. However the standards against which quality is judged holistically are implicit, being to a considerable extent dependent on the reviewer’s personal knowledge and perspective, and thus are subjective. As a result the use of implicit professional judgements as the basis for reviewing quality and identifying variations from good practice has been increasingly criticised.

Research has identified a range of assumptions about what is being measured by holistic (implicit) review and problems have been identified with the reliability and the validity of the approach. Weingart et al[6] conducted retrospective record review of 1025 case notes to compare explicit and implicit review methods when examining quality of care. Their study found that implicit reviews by physicians tend to take a global approach, including an assessment of the severity of the case, but are less likely than nurses to take into account any process issues that may lead to reduced quality of hospital care. This finding was supported by the results of a study by Gibbs and colleagues,[7] who compared quality of care for patients selected using higher and lower than expected mortality rates. The authors highlighted the insensitivity of implicit methods when used for detecting hospital level differences and reported that implicit chart reviews are not successful at discovering differences in quality of care.

Ashton et al[8] found that not only can implicit review be highly idiosyncratic and reviewer dependent, but it that it can result in lower levels of inter-rater reliability than explicit methods at patient level. Moreover, re-analysis of data from the Utah and Colorado Medical Practice Study has contributed to concerns that holistic record review may have low reliability, with the finding that different implicit review strategies produced different estimates of the total number of adverse events and negligent adverse events.[9] Despite attempts to reduce levels of subjectivity in holistic review (for example by such means as providing extensive training for physician reviewers), a number of other concerns remain about the value of review methods that are based principally on professional judgement. Inter-rater reliability between reviewers has been identified as being particularly problematic, with Hofer and colleagues finding levels of between 0.25 (low) and 0.45 (modest) in a study of a range of diseases and service settings.[10] It may also be that the choice of methods of assessing reliability may have some effect on the results of studies, since the kappa statistic is influenced by the prevalence of events.[11][12] Additionally, the individual consistency of reviewers has been questioned[13] and the individual reviewer’s bias towards harshness or leniency has been considered as problematic in comparing results between reviewers.[14] Fischoff’s initial work on hindsight bias[15] has recently been reiterated as a confounding factor in implicit review.[16] For these reasons, criterion based review, using pre-defined criteria, has been proposed as a more reliable means of assessing quality from clinical records.[17][18]

Criterion-based (explicit) methods of review are an acknowledged alternative method to holistic review and have been widely used in the UK and in the USA. Standardised methods for developing explicit evidence based review criteria were proposed by an Agency for Health Care Policy and Research working party in 1995[18] and were further developed by Hadorn and colleagues.[19] Criterion based review allows comparison of care against explicit standards (such as those derived from national clinical guidelines). It requires the definition of unambiguous questions to construct variables that can be retrieved from the case records and, although only pre-defined questions can be addressed, the variables have good reproducibility.

Derivations of these methods, using locally based standards of care as a template for identifying variations from care standards, were used in a large UK study of general practice during the 1980s.[20] Subsequently a number of structured methods for developing review criteria for explicit quality review of case notes have been developed in the UK, including methods for developing criteria directly from evidence based clinical guidelines.[21] These methods all seek to determine the rate of conformance with the criteria within a single patient’s care and are aggregated across a group of patients, recorded as a percentage. Patient preferences and clinical choices based on the severity and the anticipated outcome of the case are allowed for in estimating conformance and are not considered to be ‘violations’ of a standard of care. Criteria can thus be developed for any condition where there are externally agreed explicit standards of care.

This approach is becoming part of UK health policy. Clinical practice guidelines now being published by the National Institute for Health and Clinical Excellence are now being accompanied by evidence based review criteria to support review of clinical quality (see, for example, criteria published with the NICE guideline on the management of Chronic Heart Failure[22]).

Clinical audit in UK hospitals has adopted these objective, criterion based, approaches,[23][24][25] using explicit standards which are not profession dependent and have shown, for example, substantial variations in organisation and clinical care between hospitals.[23]

Nevertheless, criterion based review has been criticised as an insensitive method that may not identify unexpected factors that might influence outcomes of care[26][27], so that implicit review may have still retain some advantages. In some North American studies mixed methods have been adopted,[6][9] where nurses used criterion based review to identify a sub-set of problematic cases for subsequent holistic review by doctors, although this two stage approach carries a risk of hindsight bias such that those cases identified as problematic by nurses might be reviewed more harshly by the physicians.[14][17] Rubenstein and colleagues[28] proposed a structured form of implicit review in which a series of clear questions were asked of the reviewers, as distinct from seeking specific data items as in explicit review, and Pearson et al[29] used this method to review the quality of nursing care.

Thus the decision on methods is not necessarily settled by the adoption of the criterion based approach which may fail to identify the nuances of health care variation. Mohammed et al [27] reviewed the quality of care of 50 stroke patients from each of four hospitals reported as having the ‘best’ mortality outcomes for stroke in the West Midlands area of the UK and four reported as having the ‘worst’ mortality outcomes. After adjusting for case-mix using W-scores the researchers identified a number of issues that affected outcome. Some influences were predictable, such as the organisation of care. Some were unexpected, such as the influence on outcomes of ‘do not resuscitate’ orders. The authors suggested that these unexpected influences would only have been identified by expert reviewing using holistic methods.

Decisions on which of the two review methods to use, and under which circumstances, are also clouded by the results from studies that have used mixed methods. Thus Weingart and colleagues[6] have suggested that nurses and doctors may use different types of information on which to make judgements of care quality (and thus may come to different judgements about an individual case). On the other hand, Keeler et al[30] used explicit and implicit methods and sickness (risk) adjusted outcomes to review quality in different types of hospitals. They reported similar quality of care ratings for the specific hospitals when using the two methods. Any differences in quality were thought to be the result of differences in the characteristics of the hospitals rather than the result of using different methods of record review.

Overall, then, there is real lack of clarity about choice of method for case note review – which method, in whose hands and for what purpose. Building on the international evidence, this study was designed to explore these issues further.

**Study aim and research questions**

The first study aim agreed with the research commissioners was:-

To compare the validity and reliability of two methods of case note review of quality and safety of care.

Research questions were:

* Do holistic (implicit) and criterion based (explicit) methods of case note review identify the same variations in quality within the same record?
* Do holistic (implicit) and criterion based (explicit) methods of case note review identify the same variations in quality across groups of records for the same clinical condition?
* To what extent do holistic and criterion based methods of case note review provide similar results when used by reviewers from similar professional groups?
* To what extent do holistic and criterion based methods of case note review provide similar results when used by reviewers from different professional groups?
* Which method of case note review and which staff type would be most appropriate for phase two of the study (on the relationship between recorded process of care and outcomes of care)?

**Methods**

**Choice of conditions, review methods, settings and staff**

The overall research approach was to investigate holistic and criterion based case note review methods across hospitals, with a range of risk adjusted levels of 28 day mortality, using two chronic illnesses as tracer conditions. Quality of care was assessed using each of the two review methods by reviewers from three professional groups. Each case note was reviewed using both methods and by between one and four reviewers.

***Choice of clinical condition for review***

The initial research brief for this study encouraged research teams to consider using a limited range of tracer conditions, mainly concerned with the care for people with chronic conditions. Three conditions were initially proposed for the study – care for people with Chronic Obstructive Pulmonary Disease (COPD), Heart Failure and Stroke. Following discussions with the research commissioners, COPD and Heart Failure were the two conditions chosen for the study.

*Chronic Obstructive Airways Disease (COPD)*

About 10% of admissions through hospital UK Emergency Medicine Departments are for people with COPD, which has a high mortality rate 3 months after index admission. A NICE guideline with review criteria was produced at the commencement of the study.[31] An extensive set of review criteria were available from the national Royal College of Physicians (RCP) COPD audit, including a limited number that were considered predictive of survival.

*Chronic Heart Failure*

People with heart failure often have repeated episodes of hospital re-admission. To support our choice of heart failure as one of the two study conditions, we took into account the availability of an evidence based guideline,[23], together with a limited set of audit review criteria, that had recently become available from the National Institute of Clinical Excellence (NICE) and was produced by the Royal College of Physicians Clinical Effectiveness and Evaluation Unit. The guideline and review criteria also provided a basis for developing, within the study, an externally referenced set of review criteria for safety and quality assessment for heart failure management. There were no national audit data available.

***Cases for review***

There are relatively few admissions per hospital per year of new cases of heart failure or COPD and much of the diagnostic work-up is undertaken in the primary care or outpatient setting. We therefore chose to study cases of admission for an exacerbation of either of these two tracer conditions and excluded admissions for diagnostic workup.

These working definitions for data collection were:

*Definition of an exacerbation of COPD:*

“An exacerbation is a sustained worsening of the patient’s symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, cough, increased sputum production and change in sputum colour.”[31]

*Definition of an exacerbation of heart failure:*

"An exacerbation of heart failure is a sustained worsening of the patient's symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, tiredness and swelling of the feet and/or ankles."[23]

***Choosing the number of case notes for review***

In choosing the number of case notes for review we were unable to use prior hypotheses to assist in determining how many case notes would be required for the reliability studies. We considered using van Belle’s method[32] of calculating the number of events (e.g. identified opportunities for error as being event = 20 times the number of parameters) from which to assist with this calculation, but subsequently found that it was practically impossible to model the range of opportunities for error presenting in these complex care pathways. In addition it was decided that the study was more likely to find variations in care rather than identifiable adverse events and that in health care there could be very large opportunities for error. We therefore took a pragmatic decision to select approximately 50 case notes per condition per hospital. This number also fitted with the custom and practice of the Royal College of Physicians CEEu in which about 60 case notes per hospital form the basis for review in national clinical audits and this number of case notes had previously provided sufficient data for studies of inter-rater reliability.[24]

For this first phase of the study we therefore sought to obtain sets of 50 case notes from consecutive admissions for each condition in each of eight hospitals. That is, 800 case notes in total.

***Selection and recruitment of hospitals and staff***

A four stage process was used to recruit eight study hospitals in England. First, Hospital Episode Statistics[33] on 28 day mortality data for COPD and Heart Failure were accessed through the East Midlands Public Health Observatory. Hospitals were excluded from the selection process if they reported less than 200 in-patient cases per year for either condition, effectively excluding smaller or specialist acute hospitals. There were 136 hospitals in the final data set.

Second, twenty eight day mortality data for the two study conditions for each hospital was combined using simple averaging, to create an average 28 day mortality ratio for each hospital. Third, these were then ranked from the lowest mortality to the highest and the data was split into four quartiles, each of 34 hospitals. And finally, from this ranking, hospitals were randomly selected in each of the lowest and the highest mortality quartiles.

***Combinations of review methods and proposed numbers and types of staff***

In our initial research proposal we sought to create review teams in each hospital comprising two types of personnel – clinical staff (for example, nurses and staff working in clinical audit departments) and doctors in senior stages of their specialist training (medical Specialist Registrars). This choice of types of staff was made in order to test some of the assumptions in the literature[6] that medically qualified reviewers undertake holistic case note reviews differently from other personnel. Different combinations of reviewer type would review the same records to test inter-rater reliability within types of reviewer and between types of reviewer. Each case note would be reviewed twice by each reviewer, first using holistic, then criterion based, methods. We chose this sequence to reduce the bias on the holistic review results that might occur from a reviewer having previously examined the case notes to undertake criterion review.

Subsequent to these initial decisions on numbers of case notes and staff a more limited set of reviews was agreed with the research commissioning panel, because the costs of undertaking a full set of two reviews of each of 50 case notes across the eight hospitals using four reviewers per condition proved too great for the available study budget,. This second proposal still retained the ability to make comparisons between types of reviewer and review methods, albeit with a smaller number of reviewers in total. Table 1 indicates the type of reviewers and frequency of review that were proposed. The number next to the code for Clinical Audit/Nursing Staff (CA) and Physicians (P) indicates the number of reviewers sought for each condition in each hospital. Each reviewer was expected to evaluate 50 sets of case notes.

For each of the two conditions the framework in Table 1 was used to calculate the number of reviews and reviewers for both the holistic and the criterion based review methods, so that the total proposed number of case note reviews for the eight hospitals was 3600, using 400 case notes overall for each of the two conditions (800 in all).

**Table 1. Proposed numbers of reviewers and types of staff for each review method** (each staff member to undertake 50 reviews)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Upper mortality hospitals** | Hospital 1 | Hospital 2 | Hospital 3 | Hospital 4 | Total |
| **COPD** | CA (1) | CA (1) | CA (2) P(1) | CA(2) P(2) |  |
| **Heart Failure** | CA (2) P (1) | CA (1) P (2) | CA (2) | CA (1) |  |
| Total reviews (**from 400 case notes)** | 200 | 200 | 250 | 250 | 900 |
| **Lower mortality hospitals** | Hospital 5 | Hospital 6 | Hospital 7 | Hospital 8 |  |
| **COPD** | CA (1) | CA (1) | CA (2) | CA (2) |  |
| **Heart Failure** | CA (2) P (1) | CA (1) P (2) | CA (2) P (1) | CA (1) P (2) |  |
| **Total reviews** (from 400 case notes) | 200 | 200 | 250 | 250 | 900 |
| **Overall total of eviews for each review method** |  |  |  |  | 1800 |

***Recruitment of hospitals and staff***

Recruitment of study participants was a complex and time consuming process as participation required the agreement the COPD and heart failure clinical teams at each hospital and was also dependent on the availability of hospital staff to review records for the study. A total of eight hospitals were required for participation in this first part of the study. Because we expected difficulties in recruiting hospitals, mainly a lack of availability of staff within hospitals to review records, a total of 20 randomly selected hospitals (10 in the lower mortality quartile and 10 in the higher mortality quartile) were contacted and invited to participate in the study when only 8 were actually required to participate in this phase.

From the initial 20 hospitals contacted, the study recruited five hospitals in each quartile, including one reserve hospital per mortality quartile in an attempt to ensure that a minimum number of eight hospitals were available for the analysis. One reserve hospital subsequently dropped out leaving a total of 9 hospitals in the study, four in the lower mortality quartile and five in the higher mortality quartile – see Figure 2.

**Figure 2: Distribution of higher and lower study mortality hospitals**

|  |  |
| --- | --- |
| **Box plot of COPD/Heart Failure mortality showing differences between higher and lower mortality groups** | **Bar chart showing individual hospital level mortality rates** |
|  | **Hospitals with higher mortality**  **Hospitals with lower mortality** |

Hospitals were invited to participate through contact with one of the specialists in COPD and one in heart failure management. In each selected hospital, consultants specialising in each of the two conditions were approached jointly by the Royal College of Physicians Clinical Effectiveness and Evaluation Unit and The University of Sheffield with a request that they act as sponsors for the study. Their role was to recruit review staff within the hospital who would undertake the data collection for the study. Recruitment of the selected hospitals was completed when two specialists in a hospital agreed to act as sponsor and there were enough staff to undertake the reviews (Figure 3, below).

The proposed number of reviewers required for each condition at each hospital varied from one to four (see Table 1) and this request sometimes proved difficult to meet in some hospitals. Among the reasons affecting the recruitment of reviewers included whether the hospital had a dedicated audit department and the change-over time of Specialist Registrar training posts. At the end of the recruitment period, three types of hospital staff (reviewers) were engaged in the study - doctors in specialist training, other staff with a clinical background (of whom many were nurses specialising in the care of one of the two tracer conditions), and non-clinical audit staff. Across the nine participating hospitals the reviewers comprised 16 doctors, 10 specialist nurses (together with one clinically trained audit person, one pharmacist and one physiotherapist) and nine non-clinical audit staff (i.e. 38 reviewers in total).

**Figure 3**

11 hospitals do not wish participate

4 hospitals with low mortality rates (HES data)

5 hospitals with high mortality rates (HES data)

Each reviewer reviews 50 (max) COPD or HF records using holistic review method

*Note*: *Where there is more than 1 reviewer for a condition at the same site, all reviewed the same 50 records*

Each reviewer exports holistic review data to the study team for analysis

Site: A

reviewers COPD: 3

HF: 2

Site: C

reviewers COPD: 1

HF: 0

Site: D

reviewers COPD: 3

HF: 2

Site: E

reviewers COPD: 3

HF: 2

Site: F

reviewers COPD: 1

HF: 3

Site: G

reviewers COPD: 1

HF: 3

Site: H

reviewers COPD: 1

HF: 3

Site: J

reviewers COPD: 3

HF: 2

Each reviewer then reviews the same (max 50) COPD or HF records using criterion based review method, using same patient records as for holistic review

Each reviewer exports criterion review data to the study team for analysis

20 hospitals invited to take part in Phase 1

Site: B

reviewers COPD: 1

HF: 4

#### Data capture methods

***Holistic review***

The concept of structured implicit review [28] has been found to be valuable in North American implicit review studies as a means of reducing the variability previously found in inter-rater reliability studies. [10,11] Structured implicit review attempts to place a framework on data collection by providing headings that can be used in the ‘holistic story’. However, US based authors working for the RAND Corporation [28] chose to use what might be termed a mid-point between criterion based and textual holistic review, using structured questions that were not as specific as review criteria but which organised responses to the questions in such a way that might be considered to ask closed questions of the data. See, for example, Box 1.

Box 1

|  |  |
| --- | --- |
| Was the length of stay appropriate given the patient’s status at discharge and disposition plans? | |
| Definitely yes | \_\_\_\_\_1 |
| Probably yes | \_\_\_\_\_2 |
| Probably no | \_\_\_\_\_3 |
| Definitely no | \_\_\_\_\_4 |

From: Rubenstein et al 1991[28]

In this study the concept of structured holistic review was developed to provide reviewers undertaking holistic reviews with a limited structure but one that was at the same time not so directive as the structured implicit review framework developed by the RAND teams. In doing so, this allowed for different levels of health care quality to be identified, from excellent care, to care not provided, to the identification of adverse incidents.

Data was captured under three phases of care and for care overall:-

* Care during the investigation/assessment phase
* Care during the initial management phase
* Care during the pre-discharge phase
* Quality of care overall

Using this structured holistic framework, reviewers were asked to provide two forms of assessment of quality and safety of care. First, reviewers provided a written assessment of the quality and safety of care of each patient, using information from the case notes (paper and/or electronic records) of the most recent episode of in-patient care for an exacerbation of the illness.

For the phases of care reviewers were guided by two prompts:

* Please comment on the care received by the patient during this phase, and
* From the records, was there anything in particular worth noting

Second, reviewers were asked to rate the care received by the patient for each of three phases of care - admission/investigations, initial management and pre-discharge care. Each phase was rated on a 1-6 scale (1 = unsatisfactory, 6 = very best care) and a definition was provided for each of the points on the scale:-

1. Care fell short of current best practice in one or more significant areas resulting in the potential for, or actual, adverse impact on the patient

2. Care fell short of current best practice in more than one significant area, but is not considered to have the potential for adverse impact on the patient

3. Care fell short of current best practice in only one significant area but, is not considered to have the potential for adverse impact on the patient

4. This was satisfactory care, only falling short of current best practice in more than two minor areas

5. This was good care, which only fell short of current best practice in one or two minor areas

6. This was excellent care and met current best practice.

The format of the questions is set out in Box 2.

Box 2

**Investigations/examination (for example)**

We are interested in comments about the quality of care the patient received and whether it was in accordance with current best practice (for example, your professional standards). You may also wish to comment from your own professional viewpoint. If there is any other information that you think is important or relevant that you wish to comment on then please do so.

Please comment on the care received by the patient during this phase.

From the records, was there anything in particular worth noting?

Please rate the care received by the patient during this phase.

Please tick only one box

Unsatisfactory [] [] [] [] [] [] Very best care

Next, in assessing the quality of care overall, reviewers were asked:

* Please comment on the care received by the patient overall.

An overall quality of care rating was requested for each patient review on a 1-10 scale (1 = unsatisfactory, 10 = very best care, only using the two anchor points on the scale) to provide for a global rating of care quality. This was given a wider, more fine grained, scale so that reviewers could assimilate their perceptions of care for all of the phases of care to give an ‘in the round’ assessment (Box 3).

Box 3

**Overall Assessment**

Please comment on the care received by the patient overall.

Please rate the care received by the patient overall.

Please tick only one box

Unsatisfactory overall [] [] [] [] [] [] [] [] [] [] Very best care overall

***Assessing the quality of recording in the case notes***

Evaluation of the quality of care through case note review is critically dependent on the quality of recording in the case notes, together with that in associated data sources such as computerised pathology and radiology results. It might be hypothesised that a poor record could prevent a high quality retrospective critical review of care. Alternatively, there might be a relationship between poor case notes and poor quality of care. Factors enhancing the use of the record for case note review include the extent to which information is recorded and placed in the case notes, the detail or otherwise of the entry and the legibility of the entry.

It was anticipated that most of the information relevant to the study would be recorded on paper based case notes but that systems would vary from hospital to hospital, for example in the extent to which the principal case notes provided a holistic record of care or whether medical notes and nursing notes might be held separately.

Reviewers were therefore asked to assess the quality of each record at the end of the holistic review, using a six point rating scale (1 = inadequate, 6 = excellent) :-

1. The patient record contains gaps in three or more significant areas

2. The patient record contains gaps in two significant areas

3. The patient record contains gaps in one significant area

4. The patient records are satisfactory and only contain gaps in three or more minor areas

5. The patient records are good and only contains gaps in one or two minor areas

6. The patient records are excellent.

Reviewers were asked to complete their assessment in the form shown in Box 4.

Box 4

|  |
| --- |
| We are interested in your view about the quality of the patient records in enabling good quality care to be provided.  Please tick only one box  Inadequate [] [] [] [] [] [] Excellent |

***Review criteria development for COPD and Heart Failure***

Criterion based review does not seek judgements of care, only requiring the reviewer to identify and record specific items of care. The purpose of review criteria when used in clinical audit is to gather data on which to make a judgement about the quality of care provided by an institution. However, for the purpose of this study, although the quality of care provided by the hospital was useful information, the prime objective was to investigate the extent to which data collection of a case note review method was reliable and in which staff type’s hands was it most reliable.

This objective meant that the number of review criteria used for each of the two conditions could be limited to a smaller number rather than, for example, the full set used by national clinical audit projects (for instance the Royal College of Physicians COPD audit comprised about 75 clinical criteria in total).[34]

The review criteria were developed using established methods for developing explicit evidence based review criteria from clinical guidelines.[19][20][22] That is, for each of the two conditions, the first draft of the criteria were developed from the evidence base in the relevant national clinical guideline[23][31] and where then subsequently validated using expert opinion.

COPD review criteria

Information to form the first draft set of criteria came from the national clinical guideline for the management of COPD[31], the limited associated set of review criteria from the guideline and, thirdly, from the national RCP clinical audit for COPD.[34] From the guideline recommendations and the available review criteria the project team identified a subset of criteria that might be useful in the study.

Refinement of the set was undertaken in three stages. Firstly, the criteria were reviewed to determine whether the required data were likely to be available from case note review. This excluded a number of review criteria used in the national RCP audit [34] that were concerned with organisational effectiveness. Thirty eight criteria remained.

Secondly, a questionnaire was sent to a selected group of respiratory physicians to seek their views on the value of the criteria for measuring quality of care. Seventeen senior physicians and specialist nurses ranked the criteria as:

* Essential
* Desirable
* Non-essential

Eleven criteria were removed as a result of this process.

Thirdly, the structure and wording of each criterion in the dataset was reviewed to ensure that it was clear, logical and could be captured from case notes. At the end of this process there were 27 criteria for COPD care (see Appendix 1).

Heart failure review criteria

A similar approach was taken to the production of heart failure review criteria. A draft set of criteria was developed from information in the national clinical guideline for the management of heart failure [23] and from the limited associated set of criteria for the guideline. Discussion within the project team identified a subset of criteria that might be of value in the study.

Refinement of the set was undertaken in three stages. Firstly, the criteria were reviewed to determine whether the required data were likely to be available from case note review. There were 34 criteria.

Second, a questionnaire was sent to a selected group of cardiovascular physicians and specialist cardiovascular nurses to seek their views on the value of the criteria for measuring quality of care. Ten replies were received. One criterion was removed as a result of this process.

Third, the structure and wording of each criterion in the data set was reviewed to ensure that it was clear, logical and could be captured from clinical records. At the end of this process there were 33 criteria for heart failure care (see Appendix 2).

An example of the external review questionnaire used for both COPD and heart failure can be found at Appendix 3.

***Developing data capture tools***

In order to facilitate the work of the reviewers and data transfer to the study team, data capture was developed through an electronic format based on Microsoft Access©. Holistic data capture forms were developed from the format outlined in Holistic Review (above), using separate screens for key data, case history data, phases of care and overall care – see Appendix 4 for examples. The database was constructed so that information could be transferred to the study team either by email or by CD, first removing all identifiable data to preserve anonymity for patients and staff. The hospital staff retained access to the full data set to provide for local analysis and audit should they so wish. Criterion based review data collection fields were created in the same way as those for holistic data.

Because there was considerable variety of local systems and versions of Microsoft Access© in the study hospitals, copies of Microsoft Access© were purchased and made available to the reviewers where required. Provision was also made for staff to collect data on a paper form where data processing facilities were difficult to access. For these records, data entry was undertaken by the research team from anonymised paper records.

Data were collected from consecutive admissions over a period of six months before the review process started in each hospital, a time period which varied slightly but was approximately between January and July 2005.

***Reviewer training and case note selection support***

The study sought to provide all reviewers with standardised training in case note review, the emphasis here being to train in the data capture methods. Each reviewer was provided with copies of clinical guidelines for COPD and heart failure care as a means of ensuring that all reviewers had an explicit database of the standards of care expected for the two conditions.[23][31] Other than providing the guidelines, in this part of the study there was no intention to try to influence each reviewers own implicit standards for quality of care – that is, each reviewer would have their own, internal, standards for the care that they were reviewing.

During a day-long training session reviewers were provided with an introduction to the two review methods (particularly since most reviewers were not familiar with the holistic method), together with review software training. Quality of care variation was discussed using four theoretical scenarios from stroke care that contained aspects of good and poor care (see Appendix 5). Stroke care was chosen for training to avoid biasing the reviewers in their view of quality of care for the two study conditions.

The challenges of finding information in paper based records and dealing with missing data were also considered along with a discussion about the means of obtaining case notes from the hospital records departments. Particular attention was paid to identification of case notes of admissions for exacerbation of known COPD or heart failure (rather than new cases or admissions for a main condition that was not related to the study), and to selecting case notes from the most recent admission.

During the data collection period a telephone helpline was made available throughout office hours if reviewers had any queries or required advice about the data collection. The study team also contacted each reviewer regularly throughout the study period to track progress with the reviews and liaised with relevant hospital medical record departments if reviewers had problems obtaining records.

**Analysis methods**

***Overall approach***

The quantitative analysis was designed to investigate the extent of reliability between individual reviewers, and groups of reviewers of the same, and different, professional backgrounds, using measures of internal (intra-rater) consistency and between reviewer (inter-rater) reliability for holistic quality of care scale scores and criterion based scores. Correlation and regression analyses were undertaken.

Detailed qualitative analysis of the textual data provided on the phases of care and the overall care was undertaken to explore the relationship between the holistic scale scores for each case and the narrative assessment. This analysis was also used to explore any differences between the results from the different professional groups undertaking the reviews.

***Holistic scale score analysis***

To assess intra-rater consistency (that is, whether reviewers were internally consistent in their ratings of care) for each individual review, the mean scale score rating was calculated across the three phases of care (admission/investigations, initial management, and pre-discharge). The Pearson correlation coefficient was calculated between the mean rating of the three phases (each on a 1 – 6 scale) and the overall rating (on a 1 – 10 scale) within each review. The purpose of this analysis is to examine the consistency of the reviewer’s scoring across the phases of care and in the final overall care to discover, for example, whether some reviewers might have quite low scores for one or more phases of care and then a rather higher score for overall care.

Intra-class Correlation Coefficients (ICC) were used as the principal measure of agreement[35]. However, although the Kappa statistic is susceptible to prevalence [in this case, of ‘opportunity for error’ rates per set of records].[10][11] Kappa scores were also computed as measures of agreement for overall scores (see Table 6a below) since this measure of agreement is more commonly used in the literature, although kappa for these scores may be influenced by the prevalence effect.

To assess inter-rater reliability between ratings of the same record by different reviewers, raw ratings were converted to ranks to adjust for variation in the range of scores used by different reviewers and Intra-class Correlation Coefficients (ICC) were calculated on these ranks.

***Measuring reliability between reviewer pairs***

The Intra-Class Correlation coefficient (ICC) gives the correlation between any two measurements or ratings for the same subject or patient, using randomly chosen methods or reviewers. ICCs are based on continuous data, unlike kappa statistics which require the data to be categorical. ICCs were used to assess the reliability between reviews of the same patient records carried out by pairs or groups of reviewers (e.g. 2 nurses or 2 doctors) at the same hospital and were calculated firstly between the holistic quality of care ratings allocated by the two reviewers, and secondly between the criterion-based scores.

When undertaking the holistic (implicit) review, each reviewer rated the overall quality of care received by each patient against a 10-point scale. It is possible that different reviewers may have interpreted the rating scale differently (e.g. one reviewer may tend to give higher or lower ratings than another). Therefore, each reviewer’s ratings were converted to a rank. For example, if a reviewer reviewed 50 records then the ratings were ranked from 1 to 50. (In the event of tied ratings, the average rank was used.) The reliability between these ranked ratings for each pair of reviewers was then assessed using ICCs.

For the criterion-based review, care was assessed against a set of condition-specific criteria for either COPD or heart failure care. ICCs were used to assess the inter-rater reliability between the overall number of criteria (as un-ranked criteria scores) noted by each reviewer as having been met.

Average reliabilities per staff type

To provide an overview of the average reliability for each staff type (e.g. doctors versus doctors or nurses versus nurses), a pooled or overall mean ICC was calculated across all the reviewer pairs in each staff group. Because some reviewer pairs had reviewed more records than others, each ICC was weighted when calculating the overall mean ICC, with the weight being proportional to the inverse of the variance of the ICC estimate.

Sites with more than 2 reviewers of different types

For sites where there were two reviewers of one staff type plus one of another type (e.g. two doctors and one nurse), we wished to avoid counting the same nurse twice in the comparison with the doctors. Therefore, the mean of the two doctors’ scores for each record was calculated (and the mean holistic scores converted to a rank). An ICC was calculated between the mean score from the two doctors and the score from the nurse. This approach was used whenever there were odd numbers of a reviewer type in this analysis. The ICC was then combined with the doctor-nurse ICCs from other sites to calculate an overall mean ICC for doctors versus nurses, weighting by inverse variance as described above. At site B there were three rather than two doctors reviewing the same records. Therefore, a single ICC was calculated among all three doctors’ scores at this site.

Care was rated on a 3 point scale; Care fell short of current best practice (Unsatisfactory)/ Satisfactory/Good or excellent care. We considered whether it might be possible to reduce down the scale score data to a binary ‘poor’ or ‘good’ score to enable direct comparisons on a 2x2 table between the two review methods but this approach would reduce the spread of judgements even further from the 6 or 10 point scales and would not accommodate the range of judgements offered by the reviewers.

***Criterion based review***

Data were scored in two ways, firstly to assess the completeness of the data and to assess the effectiveness of each reviewer type at completing the data collection form and, secondly, to calculate a quality of care score for each review.

An ‘effectiveness of reviewer’ score for each record review comprised one point allocated for each data field completed by the reviewer (irrespective of whether the criterion was recorded as being met or recorded as not being met), and one point subtracted for every data field left blank by the reviewer. These scores were converted to a percentage.

Quality of care scores for each record comprised the percentage of the criteria identified by the reviewer as having been met. Intra-class Correlation Coefficients were used to estimate inter-rater reliability for overall scores by pairs or triplets of staff reviewing the same records. Because some phases of care only generated a small number of criteria, ICCs were not computed for phases of care.

Intra-class Correlation Coefficient estimates from the different combinations of reviewers were pooled using a weighting that was inversely proportional to the variance of the estimate.[37]

***Comparison of holistic scale scores and criterion based review***

Inter-rater reliability results for each of the two review methods were compared. Additionally, an estimate of the within-staff-type consistency across the two review methods was calculated using p-values for differences between the overall holistic quality of care ratings and the percentage of criteria recorded as being met.

***Comparison of quality scores with hospitals grouped by mortality level***

In the original call for proposals for this study it was suggested that the quality of care scores might be risk adjusted by severity of illness of each case. However we agree with Daley and colleagues[37] that risk adjustment remains a controversial and difficult subject. Pitches et al[38] undertook a systematic review of 36 studies, which included 51 ‘process-versus-risk-adjusted-mortality relationships’, exploring the extent to which variations in risk-adjusted mortality rates were associated with differences in quality of care. They found a positive correlation in only 51% of the relationships with no correlation in 31% and an unexpected correlation in a further 18%, in what was a very heterogeneous set of studies.

A range of approaches to risk adjustment were considered in the initial phases of the study. Because of the complexity of data capture and the level of workload that could reasonably be asked of the reviewers, individual scoring of the risk for individual patients proved impossible. The chosen approach was therefore to compare aggregate quality of care results between the hospitals in the low and high mortality rate groups.

Analysis of holistic textual data

The comments made by the reviewers in their holistic reviews were in 2 free text areas. The first asked them to comment on the care received during a particular phase of care or for care overall. The second asked if there was anything in particular worth noting from the records about the care (see Box 2, above).

Reviewers provided textual data when commenting on each of the four phases of care and the overall assessment of care. These data were analysed in two ways in order to address the question of whether, as suggested by Weingart and colleagues[6], different staff types were concerned with different elements of care when making their holistic assessments. Data from the past medical history heading was excluded from the analysis because there were few comments and all were about the case notes rather than about care.

Content analysis

The primary approach was a content analysis drawing on grounded theory.[39] The text responses provided insights into the different ways that different individuals and different professional groups interpreted the task as well as their interpretation of care provided. By analysing textual responses we were able to investigate similarities and differences between individuals about their interpretation of the same record but also to construct pictures of how professional groups interpreted the task, care provision and give an indication of the concepts that they used.

Categorising and coding types of comment made by reviewers

Following familiarisation with the textual responses it became clear that different types of comments were given, reflecting in large part different reviewer types. A categorisation was developed that identified these different responses, irrespective of professional background. These categories could be thought of as hierarchical to a degree, if set in the context of what might constitute an ideal review. Thus at the lower end of the hierarchy were no comments or limited comments about the record rather than the care, ranging through different types of comment about care to the higher end of the hierarchy where the most discerning reviews picked up more complex issues. At the upper end there was a clear cluster of issues commented on overall, displaying a fairly sophisticated degree of reviewing. Whilst the concepts emerged from the data, the labels attached to these categories were developed by the research team. The concept terms were not used by the reviewers.

An initial set of codes was developed by five analysts in group discussion. Two pairs of analysts then each separately reviewed output from three reviewers, one pair examining COPD comments and the other examining heart failure comments. A fifth analyst examined all of the comments. Each pair of reviewers discussed their experience of using the comments and compared their results. The results of the initial analysis and commentary on the utility of the coding framework were then discussed by the group, moderated by the fifth analyst and refinements where made to the coding frame.

These categories were then used to code all responses that made up the responses for each review. Since some of the responses were made up of a number of separate comments, the code given was the ‘highest level’ category used in each of the comments. For each phase of care, and for care overall, up to four codes were allocated by the analyst. The analysis reported here refers to the highest level of code allocated by a reviewer for overall care for each set of case notes and refers only to the data collection item “Please comment on the care the patient received”. Thirteen coding categories were developed and these were also grouped into three broader categories (Box 5).

**Box 5**

|  |  |  |
| --- | --- | --- |
| **Code** | **Highest’ level comment used in the each review** | **Broad category description** |
| **1** | Blank | Codes 1-5:  Little or no comment about care and little or no judgement |
| **2** | No comment or other words to indicate nothing to say |
| **3** | Description of what’s in the record |
| **4** | Judgement of record (not care they received) |
| **5** | Description of what happened to patient (not care they received) |
| **6** | Description of care delivered | Codes 6-8:  Limited comment about care and implied judgement |
| **7** | Description of omission of care |
| **8** | Implied judgment of care (not records or patient pathway) |
| **9** | Explicit judgement of care (not records or patient pathway) | Codes 9-13:  Sophisticated comments about care with explicit judgements and views |
| **10** | Questioning/query of care delivered |
| **11** | Explanation/justification of care delivered |
| **12** | Alternative/justification of care that should have been delivered |
| **13** | Concerns |

This categorisation was subsequently used to help identify the types of reviewing undertaken by different professional groups, which in turn assisted the decision on which group or groups of professionals best matched our requirements in the review process. Together with this categorisation of the type of reviewing being undertaken, the textual analysis was also used to identify specific issues raised by reviewers and to see if they varied by professional group and by individual for the same record. Careful examination was undertaken of the particular words, phrases and style used in each comment, although not to the level of a discourse analysis. The constant comparisons allowed us to generate categories (themes) to identify different approaches to reviewing, different content of reviews and contrasts between individuals and professional groups.

***Resource analysis***

In the initial proposal to the funding body we set out a proposal for a cost benefit analysis of the two review methods. Because of resource constraints this did not form part of the final agreement. However we decided to undertake a limited resource analysis in case choices on reviewer type might essentially be made on cost (where there were limited differences in review results between one or more types of reviewers or review methods).

The resource impact of each reviewer type was explored, based on self reported data on the time taken to undertake each review and on annual staff cost data taken from Unit costs of Health and Social Care 2005 (clinical and doctor reviewers) and mid point administrative and clerical staff costs from Whitley Council pay rates (for non-clinical audit reviewers).[40] The mid point on the scale was used as the cost for each staff type and only one cost for each staff type was used in the analysis.

Descriptive statistics for the time taken to undertake each review and the cost per review for each staff type were produced in SPSS. A mean time per review and a mean cost per review for each staff type was calculated. We also included the minimum and maximum range to look more closely at the spread of the data.

**Research ethics review and research governance**

A research ethics review of the study was sought from the Trent Multi-centre Research Ethics Committee on 21 July 2004, prior to the start of data collection. Because, in both phases, data were to be collected by staff working in each hospital, and the data were anonymised before transmission to the research team, the Committee considered this to be equivalent to a national audit programme. The Trent Multi-centre Research Ethics Committee response was therefore that the study did not require an ethics opinion from the Committee.

The potential need for NHS research governance review existed in both parts of the study. However, because the data collection was being undertaken by hospital staff, the results were available to the individual hospital and the research team were not undertaking data collection on the hospital premises, the project was seen by research governance departments as akin to the national clinical audit programmes from which learning is derived as a result of the use of anonymous, collated data. No study hospital required that a full research governance review should be undertaken, although initial discussions were held with a number of research governance teams and the offer to undertake the governance review process was made to all hospitals.

**Results**

Across the nine hospitals, 38 reviewers undertook 1473 holistic reviews and 1389 criterion based reviews (a total of 684 clinical records were reviewed). The numbers of case notes reviewed by each individual ranged from nine to 50 (Table 2)*.* Variation in the numbers of reviews achieved was related to job rotations, local workload pressures and difficulties in obtaining clinical records.

**Table 2: Summary of number of case note reviews and review staff per hospital**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **COPD** | | | **Heart Failure** | | |
| **Site** | **Review staff types** | **Holistic reviews** | **Criterion based reviews** | **Review staff types** | **Holistic reviews** | **Criterion based reviews** |
| A | Non-clinical audit | 49 | 30 | Doctor | 11 | 11 |
|  | Non-clinical audit | 49 | 44 | Non-clinical audit | 12 | 12 |
|  | Doctor | 48 | 33 |  |  |  |
| B | Non-clinical audit | 50 | 50 | Non-clinical audit | 49 | 49 |
|  |  |  |  | Doctor | 49 | 47 |
|  |  |  |  | Doctor | 49 | 46 |
|  |  |  |  | Doctor | 49 | 46 |
| C | Nurse / other clinical | 49 | 19 |  |  |  |
| D | Nurse / other clinical | 49 | 48 | Nurse / other clinical | 21 | 21 |
|  | Nurse / other clinical | 50 | 50 | Nurse / other clinical | 21 | 21 |
|  | Doctor | 34 | 34 |  |  |  |
| E | Non-clinical audit | 42 | 41 | Doctor | 14 | 14 |
|  | Non-clinical audit | 43 | 43 | Doctor | 14 | 14 |
|  | Doctor | 41 | 37 |  |  |  |
| F | Nurse / other clinical | 46 | 46 | Non-clinical audit | 9 | 10 |
|  |  |  |  | Doctor | 22 | 14 |
|  |  |  |  | Doctor | 48 | 47 |
| G | Nurse / other clinical | 35 | 35 |  |  |  |
|  | Non-clinical audit | 38 | 36 |  |  |  |
|  | Doctor | 50 | 50 |  |  |  |
|  | Doctor | 50 | 50 |  |  |  |
| H | Nurse / other clinical | 49 | 50 | Nurse / other clinical | 50 | 50 |
|  |  |  |  | Nurse / other clinical | 50 | 50 |
|  |  |  |  | Doctor | 49 | 50 |
| J | Nurse / other clinical | 30 | 29 | Nurse / other clinical | 30 | 29 |
|  | Nurse / other clinical | 49 | 29 | Doctor | 25 | 24 |
|  | Doctor | 50 | 50 |  |  |  |
| **Total** | **20 review staff** | **901** | **834** | **18 review staff** | **572** | **555** |

***Quality of case note recording***

The mean quality of case notes rating for COPD and Heart Failure were 4.3 (SD 1.2) and 4.7 (SD 0.9) respectively, on a scale of 1 – 6, indicating a reasonable overall quality of recording in the paper-based notes.

***Analysis of holistic review scale scores***

Completion rates for scale scores

Data returned by reviewers was checked for completion rates. Tables 3 and 4 show completion rates in excess of 90% for all phases of care, save for the overall phase assessment completion rate for COPD reviews by non-clinical audit staff.

**Table 3: Completion rates for COPD holistic reviews**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Admission and investigations phase**  **(%)** | **Initial management phase**  **(%)** | **Pre-discharge phase**  **(%)** | **Overall phase**  **(%)** | **Total**  **(%)** |
| **Doctors**  **(n = 273 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 269  (98.5) | 265  (97) | 267  (97.8) | 272  (99.6) | 1073  (97.4) |
| **Missing data** | 4  (1.5) | 8  (3) | 6  (2.2) | 1  (0.4) | 19  2.6 |
| **Non-clinical audit (n = 271 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 260  (96) | 261  (96.3) | 263  (97) | 227  (83.8) | 1011  (93.3) |
| **Missing data** | 11  (4) | 10  (3.7) | 8  (3) | 44  (16.2) | 73  (6.7) |
| **Clinical**  **(n = 357 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 341  (95.5) | 332  (93) | 326  (91.3) | 353  (98.9) | 1352  (94.7) |
| **Missing data** | 16  (4.5) | 25  (7) | 31  (8.7) | 4  (1.1) | 76  (5.3) |

**Table 4: Completion rates for Heart Failure holistic reviews**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Admission and Investigations phase**  **(%)** | **Initial management phase**  **(%)** | **Pre-discharge phase**  **(%)** | **Overall phase**  **(%)** | **Total**  **(%)** |
| **Doctors**  **(n = 330 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 320  (97) | 323  (98) | 300  (91) | 322  (98) | 1265  (96) |
| **Missing data** | 10  (3) | 7  (2) | 30  (9) | 8  (2) | 55  (4) |
| **Non-clinical audit (n = 70 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 69  (99) | 70  (100) | 68  (97) | 70  (100) | 277  (99) |
| **Missing data** | 1  (1) | 0  (0) | 2  (3) | 0  (0) | 3  (1) |
| **Clinical**  **(n = 180 reviews)** |  |  |  |  |  |
| **Number of completed rating scales** | 170  (99) | 172  (100) | 170  (99) | 171  (99) | 708  (99) |
| **Missing data** | 2  (1) | 0  (0) | 2  (1) | 1  (1) | 4  (1) |

Intra-rater consistency in holistic reviews

Across all three staff types there were statistically significant correlations (r>0.71, p<0.001) between the mean scale score ratings that reviewers assigned to the individual phases of care and their rating of the overall quality of care, indicating a fair to good level of intra-rater consistency in rating the quality of care using holistic review scale scores (Table 5). Reviewers appeared to be relatively consistent in the way that they scored quality of care for the phases of care in a case and then gave an overall assessment score for the episode of care.

Inter-rater reliability for holistic review

Holistic review reliability between scale score ratings of the same record by pairs of reviewers was fair within all three staff types, although it varied from one reviewer pair to another and for some pairs was very poor (Table 6). The overall weighted mean ICC was fair across all three types of reviewers, with no significant differences between staff types. Table 6a displays the same analysis using kappa statistics. The same trend occurs as in Table 6, that is, the doctor reviewers have a higher level of agreement than the other staff types, although the results for the nurse/clinical group and the non-clinical audit staff are somewhat lower than for the ICC analysis. This is likely to be due to differences in the two methods of analysis.

**Table 5: Intra-rater consistency between holistic scale score ratings for phases of care and for overall care**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Review staff type (number of review staff)** | **Number of reviews** | **Mean overall rating of quality of care (SD)** | **Mean rating of phase quality of care (based on the mean score across 3 phases of care)** | **Pearson correlation between mean rating across three phases of care, and overall rating (p-value)** |
| Doctors (16) | 593 | 7.8 (1.8) | 4.7 (0.8) | 0.77 (<0.001) |
| Nurses/other clinical (14) | 529 | 7.0 (2.0) | 4.4 (1.0) | 0.81 (< 0.001) |
| Non-clinical audit (9) | 296 | 7.9 (1.3) | 4.6 (0.8) | 0.71 (< 0.001) |

1. Overall quality of care was rated on a 1 (unsatisfactory) – 10 (very best care) scale.

2. Quality of care in each of the three phases: (admission/investigations, initial management, and pre-discharge) was rated on a 1 (unsatisfactory) – 6 (very best care) scale.

**Table 6: Inter-rater reliability (ICC) between holistic overall ratings of the same record by paired reviewers of the same staff type**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Reviewer pairs** | **Condition** | **Site a** | **No of paired reviews** | **Intra-class correlation (ICC) between ranked scores (95% CI)** | **Weighted mean ICC b (95% CI)** |
| Doctor vs Doctor | Heart Failure | B **c** | 49 | 0.67 ( 0.54 to 0.79) | 0.52  (0.41 to 0.62) |
| COPD | G | 48 | 0.33 ( 0.05 to 0.56) |
| Heart Failure | F | 18 | -0.03 (-0.48 to 0.43) |
| Heart Failure | E **d** | 12 | -0.44 (-0.80 to 0.15) |
| Nurse/clinical vs Nurse/clinical | Heart Failure | D | 21 | 0.74 ( 0.47 to 0.89) | 0.46  (0.34 to 0.59) |
| COPD | D | 49 | 0.37 ( 0.10 to 0.58) |
| COPD | J | 26 | 0.27 (-0.12 to 0.59) |
| Heart Failure | H | 48 | 0.22 (-0.07 to 0.47) |
| Non-clinical audit staff vs Non-clinical audit staff | COPD | A | 48 | 0.47 ( 0.22 to 0.66) | 0.47  (0.22 to 0.66) |

**a** Only sites with more than 1 reviewer of the same staff type are included in this table

**b** Mean ICC per staff type, weighted by inverse variances to account for differing numbers of paired reviews.

**c** A single ICC was calculated for the 3 doctors at site B

**d** The doctors at site E were non-specialist doctors

**Table 6a: Kappa agreement statistics for holistic overall ratings of the same record by paired reviewers of the same staff type**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Reviewer pairs** | **Condition** | **Site a** | **No of paired reviews** | **Kappa (95% CI)** | **Weighted mean Kappa b (95% CI)** |
| Doctor vs Doctor | Heart Failure | B **c** | 49 | 0.60 ( 0.48 to 0.72) | 0.51  (0.40 to 0.61) |
| COPD | G | 48 | 0.25 ( -0.01 to 0.51) |
| Heart Failure | F | 18 | 0.35 (-0.14 to 0.83) |
| Heart Failure | E **d** | 12 | 0.00 (-0.57 to 0.57) |
| Nurse/clinical vs Nurse/clinical | Heart Failure | D | 21 | 0.34 ( 0.02 to 0.67) | 0.22  (0.08 to 0.36) |
| COPD | D | 49 | 0.26 ( 0.03 to 0.48) |
| COPD | J | 26 | 0.16 (-0.17 to 0.48) |
| Heart Failure | H | 48 | 0.10 (-0.20 to 0.40) |
| Non-clinical audit staff vs Non-clinical audit staff | COPD | A | 48 | 0.30 ( -0.01 to 0.61) | 0.30  (-0.01to 0.61) |

**a** Only sites with more than 1 reviewer of the same staff type are included in this table

**b** Mean kappa per staff type, weighted by inverse variances to account for differing numbers of paired reviews.

**c** A single kappa was calculated for the 3 doctors at site B

**d** The doctors at site E were non-specialist doctors

Overall Care was rated on a 3 point scale; Care felt short of current best practice (Unsatisfactory)/ Satisfactory/Good or excellent care

Comparisons between professional groups

Where reviewers from different staff types used holistic scale score methods to review the same record, inter-rater reliability was assessed within and between staff groups for all phases of care and overall care (Table 7). For the phase of care findings within staff groups there was generally modest to fair agreement within pairs, particularly among the doctors, although even in this group the range is large (see, for example, the initial management results) and as we have stated in the paragraph above (relating to table 6), the level of agreement varied from one reviewer pair to another and for some pairs was very poor. However, where staff from different groups reviewed the same record, agreement between the different professional groups on their assessment of the quality of care is poor to non-existent.

The overall ‘quality of care’ score for both holistic and criterion based methods used across the 684 patient records was similarly rated by the three staff types (between 70% and 79%, where 100% is very best care). Analysis of variance between the holistic overall scale ratings of the three staff types show that the nurse/other clinical group scores were significantly lower than the doctor (p<0.001) and non-clinical audit groups (p<0.001). The comparison of the latter two groups showed no significant differences (p=0.352).

**Analysis of review criterion based scores**

Criterion based reviewer effectiveness

Effectiveness scores relate to the ability of the reviewer to find and access the data in the case record, for each criterion (one point allocated for each data field completed by the reviewer (irrespective of whether the criterion was recorded as being met or recorded as not being met), and one point subtracted for every data field left blank by the reviewer. Effectiveness in capturing criterion based data was high and similar across all three staff types (Table 8), with mean scores all around 95% (that is, an average of approximately 1.5 data items missing for each review).

Inter-rater reliability for criterion based review

Inter-rater reliability between criterion based scores (that is, the percentage of criteria recorded as being met) for the same record by different reviewers ranged from moderate to good within all staff types, although with the doctors showing a significantly higher level of reliability (Table 9).

**Table 7: Within staff type ICC results and between staff type group comparisons of inter-rater reliability of holistic scale scores for phases of care and overall care**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Reviewer pairs** | **No of reviewer pairs (or triplets)** | **No of case notes** |  | **Weighted mean ICCa between ranked scores** | | | | | |
| **Admission / Investigations & examinations** | **Initial Management** | **Pre-discharge** | | **Overall** | |
| ***Within staff type ICC results*** |  |  |  |  |  |  | |  | |
| Doctor vs Doctor | 4 | 127 | **Wtd mean:** | **0.58** | **0.70** | **0.46** | | **0.52** | |
| **95% CI:** | 0.48 to 0.68 | 0.63 to 0.78 | 0.34 to 0.59 | | 0.41 to 0.62 | |
| **Range:** | -0.41 to 0.72 | -0.31 to 0.81 | -0.01 to 0.55 | | -0.44 to 0.67 | |
| Nurse/clinical vs Nurse/clinical | 4 | 144 | **Wtd mean:** | **0.50** | **0.22** | **0.43** | | **0.46** | |
| **95% CI:** | 0.38 to 0.62 | 0.07 to 0.37 | 0.30 to 0.55 | | 0.34 to 0.59 | |
| **Range:** | 0.24 to 0.76 | -0.12 to 0.41 | -0.04 to 0.77 | | 0.22 to 0.74 | |
| Non-clinical audit staff vs Non-clinical audit staff | 2 | 87 | **Wtd mean:** | **0.35** | **0.10** | **0.39** | | **0.47** | |
| **95% CI:** | 0.16 to 0.54 | -0.10 to 0.30 | 0.21 to 0.57 | | 0.22 to 0.66 | |
| **Range:** | 0.31 to 0.38 | -0.11 to 0.27 | 0.32 to 0.45 | | 0.47 to 0.47 | |
| ***Between staff type comparisons*** |  | | | | | | | | |
| Doctor vs Nurse/clinical | 5 | 179 | **Wtd mean:** | **0.23** | **0.25** | | **0.29** | | **0.43** |
| **95% CI:** | 0.09 to 0.37 | 0.12 to 0.39 | | 0.16 to 0.43 | | 0.31 to 0.54 |
| **Range:** | 0.03 to 0.38 | 0.02 to 0.41 | | -0.21 to 0.63 | | -0.06 to 0.67 |
| Doctor vs Non-clinical audit staff | 6 | 188 | **Wtd mean :** | **-0.01** | **0.03** | | **0.25** | | **0.24** |
| **95% CI:** | -0.15 to 0.12 | -0.11 to 0.16 | | 0.12 to 0.38 | | 0.12 to 0.37 |
| **Range:** | -0.15 to 0.67 | -0.53 to 0.45 | | -0.16 to 0.71 | | -0.39 to 0.54 |
| Nurse/clinical vs Non-clinical audit staff | 1 | 34 | **Wtd mean:** | **-0.12** | **0.19** | | **0.47** | | **0.43** |
| **95% CI:** | -0.44 to 0.23 | -0.15 to 0.49 | | 0.17 to 0.70 | | 0.11 to 0.67 |
| **Range:** | -0.12 to -0.12 | 0.19 to 0.19 | | 0.47 to 0.47 | | 0.43 to 0.43 |

a Weighted mean Intra-class Correlation Coefficient: estimates from the different combinations of reviewers were pooled using a weighting that was inversely proportional to the variance of the estimate

**Table 8:** **Criterion based reviewer effectiveness scores\***

|  |  |  |  |
| --- | --- | --- | --- |
| **Review staff type**  **(no of review staff)** | **Number of reviews** | **Mean score %, SD (95% CI)** | **Range** |
| Doctor (16) | 477 | 94.9, 4.8 (93.2 to 96.5) | 74.2 to 100.0 |
| Nurse/other clinical (14) | 443 | 95.2, 4.1 (93.5 to 97.0) | 67.7 to 100.0 |
| Non-clinical audit (9) | 289 | 94.7, 5.0 (93.2 to 96.5) | 61.3 to 100.0 |
| **Total (39)** | **1209** | **95.0, 4.6 (94.0 to 95.9)** | **61.3 to 100.0** |

\* Analysis excludes patients who died

\* 95% CIs are adjusted for clustering by reviewer.

**Table 9: Inter-rater reliability between criterion based scores (proportion of criteria stated as being met) for the same record by different reviewers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Reviewer pairs** | **Condition** | **Site a** | **No of paired reviews** | **Intra-class correlation (ICC) between scores (95% CI)** | **Weighted mean ICC b (95% CI)** |
| Doctor vs Doctor | Heart Failure | F | 14 | 0.96 ( 0.87 to 0.99) | 0.88  (0.83 to 0.93) |
| COPD | G | 50 | 0.65 ( 0.46 to 0.79) |
| Heart Failure | B | 46 | 0.65 ( 0.50 to 0.77) |
| Heart Failure | E**c** | 12 | 0.64 ( 0.13 to 0.88) |
| Nurse/clinical vs Nurse/clinical | COPD | J | 25 | 0.86 ( 0.71 to 0.94) | 0.74  (0.66 to 0.82) |
| COPD | D | 48 | 0.70 ( 0.52 to 0.82) |
| Heart Failure | D | 21 | 0.69 ( 0.38 to 0.86) |
| Heart Failure | H | 50 | 0.27 ( 0.00 to 0.51) |
| Non-clinical audit staff vs Non-clinical audit staff | COPD | E | 40 | 0.69 ( 0.49 to 0.82) | 0.61  (0.47 to 0.76) |
| COPD | A | 29 | 0.33 (-0.04 to 0.61) |

**a** Only sites with more than 1 reviewer are included in reliability analysis, therefore some sites do not appear on this table

**b** Mean ICC per staff type, weighted by inverse variances to account for differing numbers of paired reviews. A single ICC was calculated for the 3 doctors at site B and this was combined with the other doctor pairs in the weighted mean ICC

**c** Non-specialist doctors

**Comparison of holistic and criterion based methods**

Table 10 shows the results of a comparison between holistic review and criterion based review methods, using ‘quality of care’ scores. Reviewers rated the overall quality of care on a 10-point scale from 1 (unsatisfactory) to 10 (very best care). This was converted to a percentage for comparison with criterion based review data. Criterion based quality of care scores are shown as percentages out of 32 criteria (where patient is a current or ex-smoker) or out of 31 criteria (where patient is a non-smoker).

Mean overall quality of care scores were similar for both holistic and criterion based methods and also for all three staff types (scores of between 70% and 79% where 100% is excellent care).

Paired individual data was used for the comparison: that is the score for each criterion review of a case note minus the overall score for the holistic review of the case note. There were 1109 paired sets of case note reviews in total (some reviewers only undertook one type of review on some case notes) so the paired review numbers are smaller than the possible total of 1384 reviews. For the purposes of the analysis there are 1109 differences (criterion review score minus overall holistic review score). The confidence intervals and p-values are adjusted for clustering by reviewer (in this case, 38 reviewers).

Estimation of the level of quality of care score agreement between the two methods for an individual record, using p-value for difference, shows that there was no significant difference between the holistic and criterion based assessments when used by the doctors (p-value for difference 0.406) and by the non-clinical audit staff (p-value for difference 0.223).

However there was a difference (that is lack of agreement) between the two scores rated by the nurse/other clinical group of reviewers. It is possible that this is because of the differences between criterion based methods and holistic methods of review. The review criteria tended to be clinical measurement based, certainly in the admission and initial management phases, whereas the qualitative data from holistic reviews (see below) suggests that the nursing trained reviewers were quite strongly influenced by the quality and effectiveness of care pathways. Holistic review results may therefore be demonstrating a view which is more nursing focussed than do the selected review criteria.

**Table 10: Mean ratings/scores of overall quality of care: paired comparison of two review methods**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Staff type** | **No of holistic and criterion based reviews1 (and review staff)** | **Criterion based review mean score as a percentage of total criteria2**  **(95% CI)** | **Holistic mean rating of overall quality of care3 (95% CI)** | **Mean difference (95% CI)** | **P-value for difference** |
| Doctor | 462 (16) | 78.7 (77.1 to 80.4) | 76.8 ( 72.2 to 81.4) | -1.9 (-6.7 to 2.9) | 0.406 |
| Nurse / other clinical | 428 (14) | 77.5 (75.0 to 80.1) | 71.2 (66.4 to 76.0) | -6.3 (-10.5 to –2.2) | 0.005 |
| Non-clinical audit | 219 (8) | 75.4 (71.1 to 79.7) | 78.5 (74.7 to 82.3) | 3.1 (-2.4 to 8.5) | 0.223 |
| All staff | 1109 (38) | 77.6 (76.2 to 79.0) | 75.0 (72.3 to 77.6) | -2.6 (-5.4 to 0.1) | 0.057 |

1. Only paired reviews are included in the analysis – that is, holistic and criterion review undertaken by the same reviewer on the same record

2. Scores are shown as percentages out of 32 criteria (where patient is a current or ex-smoker) or out of 31 criteria (where patient is a non-smoker)

3. Reviewers rated the overall quality of care on a 10-point scale from 1 (unsatisfactory) to 10 (very best care). This was converted to a percentage for comparison with criterion based review data

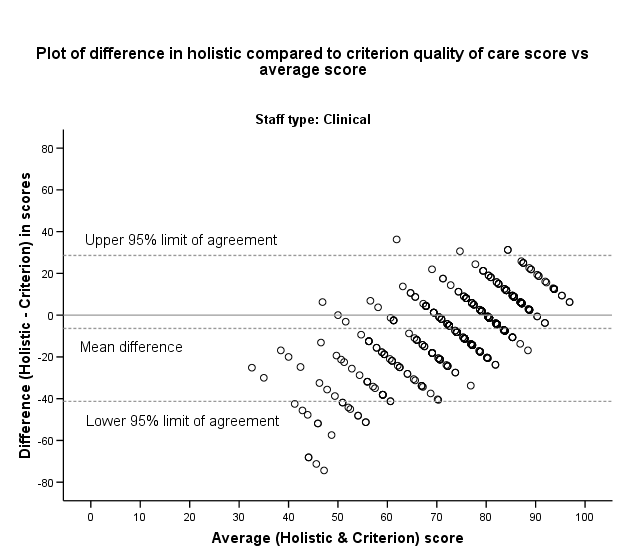
Bland – Altman plots of the difference in score between the review methods, against the average of the two scores, can also be used to examine the size of the differences and also their distribution around zero. The plot also allows for a visual check to determine whether the differences are (or are not) related to the size of the measurement.

For the purpose of this study, the average reviewer score across both methods acts as the best estimate of the true value. The mean difference, in review method scores, is an estimate of the average bias of one method relative to another. The standard deviation of the differences; or the 95% limits of agreement, can be used to see how well methods are likely to agree for an individual. For a systematic distribution we expect the range (mean +/- 2SDDIFFERENCE) to include about 95% of the observations. This range of values defines the 95% limits of agreement.

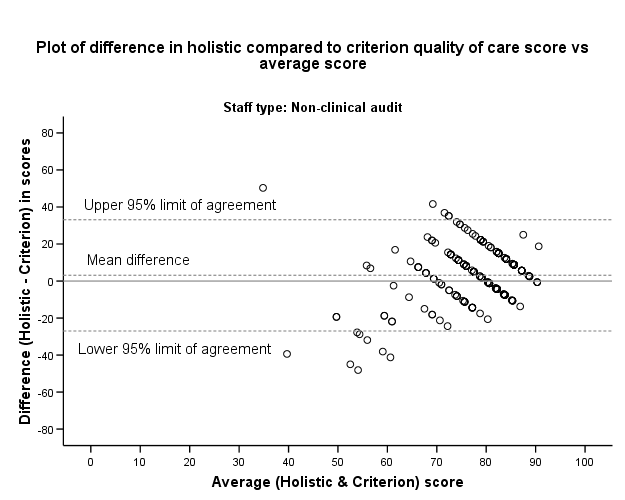
Figures 5 to 7 show that the reviewers tended to rate the majority of records with a mean combined holistic and criterion score of between 40 and 100, with very few records having lower scores reflecting poorer care. There is some evidence of a systematic pattern, in all three plots, which suggests that at lower average scores (up to 60) the holistic based score tends to be less than the criterion based score for the same reviewer/patient (i.e. negative differences). At higher average scores, above 60, then there is evidence of the opposite pattern i.e. positive differences, which implies that the holistic based scores are larger than the criterion based review scores for the same patient. These patterns, at both the higher and lower level scores, may be reflecting the methodological differences of measuring quality of care. While the criterion based scores are rigid – the item is either present or absent, holistic scoring allows the reviewer to make a judgement, which might be ‘harsher’ than criterion based scoring at lower quality levels and more ‘favourable’ at higher quality levels, since many more factors may be taken into account in the judgement in holistic reviewing. Evidence from the following section on the analysis of textual data may also support this hypothesis.

Both Table 10 and Figures 5 to 7 suggest that there is more variation in holistic review results than in criterion based review. The reasons for this may be due to the differing nature of the two review methods, since criterion based review is by its nature very structured.

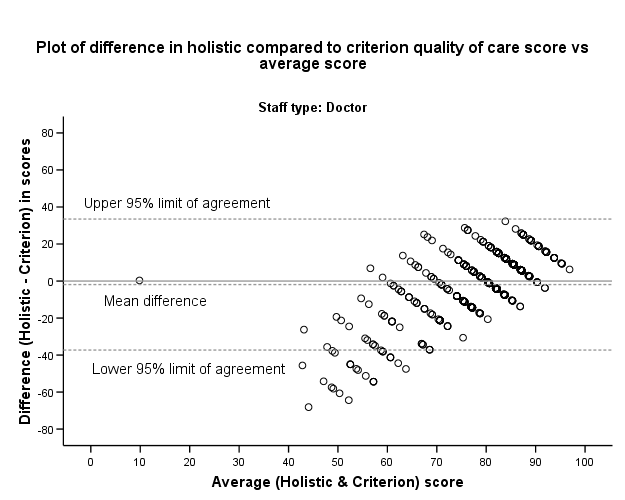
**Figure 5**



**Figure 6**



**Figure 7**



**Thematic analysis of holistic textual data**

Textual comments on the quality of care were sought from reviewers as part of the holistic review process and a textual analysis was undertaken where any type of response was given to either of the questions:

* Please comment on the care received by the patient during this phase (the first box of the data collection form)
* From the records, was there anything in particular worth noting? (the second box of the data collection form).

There was variation in the type and amount of comments given by reviewers. Some reviewers gave no response, others gave one or two words, while a list type response was given by some and extensive narratives were provided by other reviewers. The data are presented by staff type, followed by an overall summary analysis.

## **Non-clinical audit staff**

Several non-clinical audit staff reviewers made no comment in the first box in most instances. In a few instances those who made no comment did however offer a comment in the more general comments box (the second box). Among this reviewer group the comments in the second box were sometimes about documentation rather than about the care delivered.

A relatively common approach by non-clinical audit staff was to present a list, of things that had been done, or requested or in some instances that had not been done. In some instances the lists were quite exhaustive; however it was not possible to say whether the reviewer had included them because they had decided they were relevant to judgements about the care the patient received or simply because they looked important. It is difficult therefore to reach firm conclusions about how much selection had been exercised.

This attempt to reach conclusions was made more difficult by the fact that in some instances the comments were clearly lifted verbatim from the notes or, in other instances, a paraphrase was given which might perhaps include some interpretation of what had taken place. A list of items by itself without any information about why they are important or how they related to other aspects of their care means that there is not a sufficient narrative to allow a picture to be formed of what the reviewer thinks about the issue, let alone what their view is of the quality of care was delivered.

Another approach that was observed in the comments from this group was the attention to very specific issues, such as timing of medication issues, lack of follow up in terms of test results, details of transfers, information given to patient and family, timing of interventions/other aspects of care including delay length, if it occurred. This was sometimes the main approach taken. In other reviews it was combined with the list approach. Two reviewers did, however, pick up issues of adverse incidents or queried the care delivered in a couple of instances, but this was very much the exception rather than the rule.

On the whole, amongst this staff group, there was relatively little judgement directly expressed by the non-clinical reviewers about the care perceived. There was a limited amount of explicit and implied judgement but from only relatively few of the reviewers. It was unclear in some instances how much selection had been made about what to include in the comments provided.

## **Nursing and other clinical staff**

As with all staff groups, there was variation in the types of review comments made by nurses, although many reviewers in this group included an element of listing of what had, and had not, been done. In some instances there was only a list of actions and omissions. However for most cases there was also other information and in many instances implied or explicit judgement for care in general or for specific aspects of care. For example, many reviews gave information about what was done and commented on whether this was appropriate.

As with the non-clinical audit reviewers some of the nurse reviewers mentioned documentation, but it was quite clear that they more often were concerned with care issues rather than documentation. Even the concentration of some nurse reviewers on care plans (which might be expected given the importance of this in nursing care) appeared to be more concerned with the content of the plan rather than whether it was legible, which again tends to suggest a focus on care rather than documentation. In terms of comments about the plan, in most instances there was a judgement about the quality of the plan. Most of the comments about the plan were explicit judgements about it and in many instances the views about the plan were the most explicit comments in terms of judgements made. In aspects of care other than the plan, the judgements were more often implied.

In terms of the areas of care that comments were made on, as might be expected many were about areas that might be considered the responsibility of nursing staff. For example comments on review including observations, timing of medication administration, involvement of other nursing staff (such as specialist nurses), discharge planning, social circumstances, patient education and nutrition. However, a couple of nurse reviewers took a wider view and made comments on appropriateness of medication, investigations and teams involved in care. In some instances they also queried care provided. Several reviewers also picked up issues of concern including potential and actual adverse events.

Overall, the nurse reviewers commented much more on the care delivered rather than the documentation in contrast to the non-clinical audit reviewers. They also utilised the list approach and in some instances it was combined with either implicit or explicit views about care delivered. Whilst many focussed on areas more traditionally thought of as nursing realms, one or two did give views (implicit and in one case usually explicit) about medical as well as nursing care. It was easier to get an overall impression of the care delivered, what was delivered and views about the quality of care than with the non-clinical audit reviewers. However the pictures were still patchy in most instances and relied in many instances on the reader going with the implications, rather than being given an explicit view.

## **Doctors**

As with the other two staff groups, there was considerable variation in the reviewing style and comments given by different medical staff who undertook reviews. The variation in comments provided ranged from no opinion or very brief opinions of care delivered to several lines that allowed a reasonable picture of the episode of care to be gleaned.

As with other staff groups some element of listing what was done was evident in the comments from the medical staff. However where this was done the items listed were usually much fewer in number. Several reviewers gave limited comments but explicit judgements in almost all cases, so that their comments, as well as listing items, also contain some explicit judgements. If aspects of care had been omitted or delayed or the implication was that care could have been better, medical reviewers often gave further details that almost justified their implied criticism.

One of the most striking differences between this staff group and the other two groups is that almost all of the medical reviewers who gave fuller answers routinely gave explicit opinions, views or judgements about the care delivered. Medical reviewers picked up issues where the care given was queried, alternatives suggested and in some instances adverse incidents or practice said to be unsafe were picked up. These were more frequent in reviews by medical staff than in nursing/clinical groups and non-clinical staff. It was noted however that in instances where care was perhaps not as good as it should be, the reviewers were sometimes less willing to give an explicit judgement, certainly not without adding additional comments (and much more so than when care was deemed satisfactory). In instances where poor care was commented on, the comment often included a statement about it not being clear or mentioned if certain actions/treatments were done. This may, or may not, be a reflection of professional reluctance to criticise.

Overall the medical reviewers gave explicit views about the care provided, often supplemented with comments that allowed a better picture of the episode of care to be gleaned. Their focus was on medical care rather than on nursing or patient-centred issues, which might be expected, and might be considered that the comments were on the domains/items that could have a greater impact on patient outcomes.

## **All reviewers**

There was variation in the type and amounts of comments given by reviewers. Some reviewers gave no response, others gave one or two words, while a list type response was given by some and extensive narratives given by other reviewers. This variation might reflect different levels of understanding of care received, willingness to offer views about the care received, or both.

On the whole, it was most difficult to get a view of care from the non-clinical audit staff. From the nursing/clinical group, more information about care was given and generally more judgements were given, although they were often implicit and in the area of nursing care rather than care overall. The medical reviewers, whilst they tended to focus on medical aspects of care, usually gave an explicit view about the care given and picked up issues that were likely to have an impact on patient outcomes.

Except for the information about explanation and justification of actions, proposed alternatives and views about likely impact which was almost exclusively provided by the medical reviewers, the additional detail from lists provided by the other types of reviewers did little to help build a picture of the episode of care. In each group there was a variation in reviewing quality. The best in each group picked up issues of adverse incidents, unsafe care and gave a good account of the care episode including their judgement about how good it was. In general the reviews provided by doctor reviewers gave a better representation of the care provided than did those of the nursing/clinical group which, in turn, was a better representation than that of the reviews from the non-clinical audit staff.

***Analysis of the type and level of comment used by staff groups***

Comments in all of the responses for overall care were coded using the framework provided in Box 5 (page 27). The codes were then grouped into 3 main bands to provide an overall assessment of the similarity of types and levels of coding between the three professional groups. These groupings were:-

* Codes 1-5 – Little or no comment about care and little or no judgement
* Codes 6-8 – Limited comment about care and implied judgement
* Codes 9-13 – explicit judgement of care or more sophisticated comments about care with explicit judgements and views

The data are presented in tabular form in Tables 11 (COPD) and 12 (heart failure) and the analysis takes a null hypothesis. That is, there is no association between the level of coding given and staff type (i.e. the rows and columns are independent).

For the COPD reviews, across all of the phases of care and overall care, for both conditions, there were statistically significant differences between the types of comments made by the three types of staff for COPD, with the medical staff using the highest level codes across all phases of care and overall care (Table 11). There is therefore some evidence to suggest that doctors are more likely to use explicit codes than the other staff groups.

**Table 11: Types of comment made by different staff types for COPD (“highest level” comment per review)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Admission** | | | **Initial management** | | | **Pre-discharge** | | | **Overall** | | |
| **Coding bands** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** |
| Codes 1-5 (Little or no comment about care and little of no judgement) | 5% | 14% | 41% | 1% | 12% | 38% | 4% | 13% | 47% | 4% | 16% | 47% |
| Codes 6-8 (Limited comment about care and implied judgement) | 15% | 32% | 56% | 15% | 36% | 48% | 27% | 53% | 48% | 4% | 15% | 7% |
| Code 9-13 (explicit judgement of care [9] or more sophisticated comments about care with explicit judgements and views) | 81% | 54% | 4% | 84% | 52% | 13% | 68% | 34% | 6% | 93% | 69% | 46% |
| **Number of reviews (and reviewers):** | 258 (6) | 357 (8) | 271 (7) | 273 (6) | 356 (8) | 271 (7) | 273 (6) | 355 (8) | 271 (7) | 273 (6) | 352 (8) | 267 (7) |
| **Chi-square:** | 338.0 (p < 0.001) | | | 306.2 (p < 0.001) | | | 313.1 (p < 0.001) | | | 190.3 (p < 0.001) | | |

For heart failure reviews, while there are still statistically significant differences between the three groups, the differences between the groups are less strong, with similarities the doctors and nurses for phases of care and little difference between the groups for overall care (Table 12).

**Table 12: Heart Failure: Types of comment made by different staff types (“highest level” comment per review)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Admission** | | | **Initial management** | | | **Pre-discharge** | | | **Overall** | | |
| **Coding bands** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** | **Doctor** | **Clinical (nurse or other)** | **Non-clinical audit** |
| Codes 1-5 (Little or no comment about care and little of no judgement) | 10% | 5% | 6% | 12% | 4% | 1% | 19% | 3% | 17% | 10% | 3% | 4% |
| Codes 6-8 (Limited comment about care and implied judgement) | 6% | 12% | 77% | 6% | 12% | 72% | 21% | 51% | 64% | 2% | 8% |  |
| Codes 9-13  (explicit judgement of care [9] or more sophisticated comments about care with explicit judgements and views) | 84% | 83% | 17% | 82% | 84% | 26% | 60% | 42% | 19% | 89% | 89% | 96% |
| **Number of reviews (and reviewers):** | 329 (10) | 172 (5) | 70 (3) | 329 (10) | 172 (5) | 69 (3) | 324 (10) | 172 (5) | 70 (3) | 326 (10) | 172 (5) | 70 (3) |
| **Chi-square:** | 217.9 (p < 0.001) | | | 200.2 (p < 0.001) | | | 81.0 (p < 0.001) | | | 25.1 (p < 0.001) | | |

The high percentage of high-level codes in the overall section for non-clinical audit staff may be due to the fact that there were only 3 non-clinical audit reviewers for heart failure, and 49 out of 70 of these reviews were done by one reviewer. Unusually for non-clinical audit reviewers, this reviewer, in the overall phase, usually put a comment like “good all round care” or “good nursing and medical care”, hence it has been coded 9 - explicit judgement. However this judgement by the reviewer wasn’t reflected in the individual phases of care. It also appears that the nurses gave more high-level comments for heart failure than did those for COPD, perhaps because there were a greater proportion of specialist nurses in the heart failure group. Figures 8 and 9 provide a graphical representation of the data.

**Figure 8:** **COPD coding according to types of comment**

N = 273 reviews by 6 doctors

271 reviews by 7 non-clinical staff

357 reviews by 8 clinical staff





F**igure 9: Heart Failure coding according to types of comment**

|  |  |
| --- | --- |
|  |  |
|  |  |
| N = 329 reviews by 10 doctors  70 reviews by non-clinical audit staff  180 reviews by nurses/other clinical staff |  |

N = 329 reviews by 10 doctors

70 reviews by 3 non-clinical staff

180 reviews by 6 clinical staff

**Resource implications**

Tables 13 and 14 summarise the data on resource use on holistic review and criterion based review respectively. For COPD holistic review (Table 13), doctors and nurses both took a similar amount of time to review each record (approx 18 minutes per review). However the non-clinical audit staff took much longer to review each COPD record using holistic methods, with a mean review time of 34 minutes. The mean total cost per review data indicates that nurses had the lowest cost per COPD holistic review at £6.73, non-clinical staff had a cost per review of £8.53 and doctors incurred the highest cost per review at £10.16, despite taking the least time to undertake the reviews. Doctors incur higher costs per review than the other staff types because of the higher overall staff costs compared with the other staff groups in the study.

There was less variation between the different staff groups for the time taken to review heart failure records holistically (mean range 24.06 – 29.48 minutes). Doctors were again found to be the most expensive staff group, with non-clinical audit staff incurring the least cost per review.

Cost per review for Heart Failure holistic reviews were slightly lower for each staff group than for the COPD holistic reviews.

For the COPD criterion based reviews (Table 14) the non-clinical staff took the most time to complete each review. However, as with the holistic review data the doctors incurred the highest cost per review. This pattern is repeated in the Heart Failure criterion data.

**Table 13: Resource used during the holistic review process, by staff type**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Condition** | **Staff type** | **No. valid reviews**  **Missing data ()** | **Mean review time in minutes**  **(SD)** | **SD time per review** | **Time per review**  **Min - Max** | **Mean total cost per review**  **(£)** | **SD**  **cost per review**  **(£)** | **Cost per review**  **Min-Max**  **(£)** |
| COPD | Doctor | 266  (7) | 18.5 | 9.9 | 5 -52 | 10.16 | 5.46 | 2.75 – 28.60 |
| COPD | Nurse/ Other Clinical | 347  (10) | 18.6 | 9.3 | 2 - 60 | 6.73 | 3.62 | 0.67 – 27.00 |
| COPD | Non-clinical | 268  (3) | 34.12 | 16.7 | 5 - 105 | 8.53 | 4.17 | 1.25 – 26.25 |
|  | *All reviews* | *881*  *(20)* |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Heart Failure | Doctors | 309  (21) | 24.1 | 10 | 5 - 50 | 12.96 | 5.84 | 2.75 – 24.75 |
| Heart Failure | Nurses/ Other Clinical | 177  (3) | 29 | 21 | 10 - 180 | 10.54 | 7.80 | 3.33 – 60.00 |
| Heart Failure | Non-clinical | 63  (7) | 27.1 | 9.5 | 10 - 50 | 6.78 | 2.37 | 2.50 – 12.50 |
|  | *All reviews* | *493*  *(27)* |  |  |  |  |  |  |

T**able 14: Resource used during the criterion review process, by staff type**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Conditi-on** | **Staff type** | **No. valid reviews**  **Missing data ()** | **Mean review time in minutes**  **(SD)** | **SD time per review** | **Time per review**  **Min - Max** | **Mean total cost per review**  **(£)** | **SD**  **cost per review**  **(£)** | **Cost per review**  **Min-Max**  **(£)** |
| COPD | Doctors | 254  (0) | 19.5 | 7.2 | 6 - 50 | 9.16 | 5.21 | 2.50 – 32.50 |
| COPD | Nurses/ Other Clinical | 335  (1) | 15.4 | 5.8 | 3 - 45 | 5.57 | 2.16 | 1.00 – 15.00 |
| COPD | Non-clinical | 243  (1) | 30 | 26.2 | 10 - 280 | 6.00 | 2.52 | 1.33 – 18.75 |
|  | *All reviews* | *832*  *(2)* |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Heart Failure | Doctors | 311  (2) | 20.8 | 9.9 | 5 - 50 | 10.44 | 4.89 | 2.58 – 21.67 |
| Heart Failure | Nurses/ Other Clinical | 177  (2) | 20.6 | 8.1 | 10 - 60 | 6.27 | 2.79 | 2.83 – 17.00 |
| Heart Failure | Non-clinical | 71  (0) | 27.5 | 12 | 10 - 50 | 6.88 | 2.99 | 2.50 – 12.50 |
|  | *All reviews* | *559*  *(4)* |  |  |  |  |  |  |

***Quality of care – hospital mortality group relationship***

Of the nine study hospitals, five were grouped together in the lower mortality group and four were in the higher mortality group, based on a calculation from Hospital Episode Statistics data (Figure 2, page 15). This analysis explores the relationship between these two groups of hospitals, which are ranked and grouped by mortality rates, and the quality of care data expressed as the group mean holistic scale scores for overall care and the group mean percentage criterion score. Data are presented in tabular and graphical form.

Table 15 shows the mean difference, for each condition, between the holistic overall quality of care rating (based on a ten point scale) for hospitals classified as belonging to a higher mortality group or a lower mortality group. Mean difference is the mean score for higher mortality hospitals minus the mean score for lower mortality hospitals. A negative mean difference in the table indicates that the lower mortality group mean score is higher than the mean score for the higher mortality group.

The result trends are different for COPD and for heart failure. For COPD, quality of care scores tend to be higher in the lower mortality group of hospitals than in the higher mortality group and there is a significant difference in the quality scores given by the doctors. The significant score for all three groups of staff (p=0.033) is likely to be driven by the significant score for the doctors groups (p=0.012). These findings might be thought of as being in an expected direction although research shows that process/outcome relationships are by no means straightforward.[38]

For heart failure, there are smaller mean differences in a negative direction, indicating that quality of care scores are higher in the higher mortality hospital group. However the differences are quite small and none of the differences are significant.

It is difficult to interpret the findings in Table 15 overall – there is certainly no overall trend to higher quality scores in lower mortality hospitals. It may be that the contrasts are an example of the unexpected findings found by other researchers (for example, Gibbs et al,[7] where higher quality of care scores have been found among patients with poorer outcomes.

The holistic overall care review data in Table 15 are also presented as box plots in Figures 8 and 9 to demonstrate the distribution of scores from individual reviewers (identified in the tails of the distributions by anonymous review numbers), together with medians and inter-quartile ranges, for the two hospital mortality groups.

**Table 15: Relationship between the holistic overall quality of care rating (10 point scale) and mortality group**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Staff type** | **Condition** | **Higher mortality (n=4)**  **Number of reviews; mean score (SD)** | **Lower mortality (n=5)**  **Number of reviews; mean score (SD)** | **Mean difference1** | **95% confidence interval of the difference** | **P-value** |
| Doctors | COPD | 190; 7.19 (2.01) | 82; 8.55 (0.788) | 1.35 | 0.45 to 2.26 | 0.012\* |
| Non-clinical | COPD | 80; 7.46 (1.40) | 147; 7.92 (1.44) | 0.46 | -0.17 to 1.08 | 0.118 |
| Clinical | COPD | 206; 6.48 (2.24) | 147; 7.23 (1.74) | 0.76 | -0.79 to 2.30 | 0.286 |
| *All staff* | *COPD* | *476; 6.93 (2.06)* | *376; 7.79 (1.54)* | *0.86* | *0.08 to 1.64* | *0.033\** |
|  |  |  |  |  |  |  |
| Doctors | Heart  Failure | 166; 8.11 (1.94) | 156: 7.69 (1.42) | -0.42 | -1.52 to 0.69 | 0.415 |
| Non-clinical | Heart  Failure | 9; 8.22  (0.83) | 61; 8.15 (0.51) | -0.07 | -1.09 to 0.94 | 0.781 |
| Clinical | Heart  Failure | 137; 7.53 (1.56) | 42: 7.09 (2.02) | -0.43 | -1.47 to 0.60 | 0.332 |
| *All staff* | *Heart*  *Failure* | *312; 7.86 (1.78)* | *259; 7.70 (1.42)* | *-0.15* | *-0.83* *to 0.52* | *0.637* |

1. Mean difference is the mean score for the lower mortality hospitals group minus the mean score for the higher mortality hospitals group

\* P-values and 95% confidence intervals adjusted for potential clustering by reviewer.

**Figure 8: COPD** - **holistic overall quality of care box plots to show the**

**difference between the overall quality of care ratings** **for hospitals**

**classified as in a high mortality group or in a low mortality group**

|  |  |
| --- | --- |
| **Doctors** | **Clinical/Other** |
|  |  |
| **Non-Clinical Audit** | **All staff** |
|  |  |

Note: - ‘tails’ of the plots show the results from individual reviewers.

**Figure 9: Heart Failure holistic overall quality of care box plots to show the difference between the overall quality of care ratings** **for hospitals classified as in a high mortality group or in a low mortality group**.

|  |  |
| --- | --- |
| **Doctors** | **Clinical/other** |
|  |  |
| **Non-Clinical audit** | **All staff** |
|  |  |

Note: - ‘tails’ of the plots show the results from individual reviewers.

Table 16 shows the mean difference for the mean of the total percentage scores from the criterion based reviews for each condition for the higher and lower hospital mortality groups. Here there are not dissimilar findings from Table 15, with many of the criterion scores tending to be higher in the lower mortality group of hospitals. In this analysis there is also only one significant difference found but this time it is in the physician reviewers score for heart failure – a higher quality score for the lower mortality hospital group.

Overall, there does tend to be a higher quality score for criterion review in the lower mortality hospitals. However, note that when contrasting the results of Tables 15 & 16, the reviewers were using implicit judgements to score care in Table 15 but there were explicit standards set through the review criteria in Table 16. Judgement based holistic scale scoring tends to show larger ‘tails’ at both ends of scoring than does criterion scoring, which is unlikely to be influenced by outcome (Figures 8 – 11). For ‘poor’ care at the individual case level, reviewers using holistic methods can be very critical, as some of our textual data shows, and the reverse is the casefor good care**.**

The criterion based review data that are presented in Table 16 are also shown in Figures 10 and 11.

**Table 16: Relationship between the mean criterion score (scaled to 100) and mortality group**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Staff type** | **Condition** | **Higher mortality (n=4)**  **N; mean (SD)** | **Lower mortality (n=5)**  **N; mean (SD)** | **Mean difference1** | **95% confidence interval of the difference** | **P-value** |
| Doctors | COPD | 187; 77.33 (10.47) | 67; 77.99 (7.85) | 0.67 | -5.63 to 6.97 | 0.796 |
| Non-clinical | COPD | 120; 74.26 (12.90) | 124; 74.78 (9.31) | 0.52 | -9.09 to 10.13 | 0.895 |
| Clinical | COPD | 189; 73.63 (11.98) | 147; 76.02 (9.73) | 2.39 | -5.54 to 10.33 | 0.499 |
| *All staff* | *COPD* | *496; 75.18 (11.77)* | *338; 75.96 (9.28)* | *0.78* | *-3.27* *to 4.83* | *0.691* |
|  |  |  |  |  |  |  |
| Doctors | Heart  Failure | 163; 71.41 (14.65) | 150; 77.90 (9.23) | *6.49* | 1.52 to 11.46 | *0.016\** |
| Non-clinical | Heart  Failure | 10; 79.00 (7.74) | 61; 74.96 (8.72) | *-4.04* | -20.39 to 12.31 | *0.399* |
| Clinical | Heart  Failure | 135; 80.0 (5.62) | 42; 80.65 3.86) | *0.65* | -2.98 to 4.27 | *0.664* |
| *All staff* | *Heart*  *Failure* | *308; 75.41 12.12)* | *253; 77.64 (8.62)* | *2.23* | *-2.55 to 7.00* | *0.340* |

1. Mean difference is the mean score for the lower mortality hospitals group minus the mean score for the higher mortality hospitals group

\* P-values and 95% confidence intervals adjusted for potential clustering by reviewer.

**Figure 10: COPD criterion based total score box plots to show the difference between the criterion based total score** **for hospitals in the higher mortality group or lower mortality group**

|  |  |
| --- | --- |
| **Doctors** | **Clinical/other** |
|  |  |
| **Non-Clinical Audit** | **All staff** |
|  |  |

Note: - ‘tails’ of the plots show the results from individual reviewers.

**Figure 11: Heart Failure criterion based total score box plots to show the difference between the criterion based total score for hospitals in the higher mortality group or lower mortality group**

|  |  |
| --- | --- |
| **Doctors** | **Clinical/other** |
|  |  |
| **Non-Clinical Audit** | **All staff** |
|  |  |

Note: - ‘tails’ of the plots show the results from individual reviewers.

**Feedback to study hospitals and teams**

Following the completion of the reviews, each reviewer received a feedback report containing there own data. Individual hospitals or reviewers were not identified in the report. The report contained summary statistics and frequencies for the algorithmic criterion and the holistic quality of care scores for all phases of care and the quality of records element of the review. The holistic quality of care scores were compared with those from other reviewers reviewing the same records. A mean quality of care score was calculated for each hospital and this was also provided in the report. Appendix 6 contains an example report that has been anonymised.

Each reviewer received an electronic copy of the report since it was envisaged that the report would provide a useful basis for local audit presentations. Where requested, individual assistance was provided to reviewers wishing to present their data at local audit meetings. For example, advice was given on the meaning of the data or producing specific graphs/tables in PowerPoint.

From the feedback we received from the reviewers about the reports, most found the reports to be useful and interesting.

**Conclusions**

Retrospective assessment of the quality and safety of care primarily depends on review of information from the clinical record and the literature suggests that both holistic and criterion based review methods make valuable contributions but also have methodological limitations.

This is the first study in the UK that has compared the two methods of review and it has additionally contrasted the results of three different professional groups. Few international studies have contrasted the review results of different professional groups using the same methods, although Weingart et al[6] compared the results of explicit [criterion based] review undertaken by nurses with implicit review of the same record undertaken by physicians. They found that when examining medical records, ‘nurse and physician reviewers often came to substantially different conclusions’.[6]

Key results from the study include:

* Reviewers are relatively internally consistent
* The is some evidence of moderate to good within group reliability for holistic review
* All three professional groups were good at criterion review
* There is evidence of agreement between the results of holistic and criterion review in the hands of physician reviewers
* The is some difference in review focus between nurses and physicians

Our most important research questions have been first, to determine the level of agreement between healthcare professionals, from different backgrounds, when they review the same record. This agreement, or reliability, relates to the repeatability of the results from the review – whether a different reviewer would come to the same conclusion about the quality of care from the same data source, using the same method. This is clearly a practical question, as well as a research question. Second, we have used review of the same record to explore the relationship between holistic and criterion based methods.

Reviewers undertaking holistic review, using scale scores, were relatively consistent in the scores for each case note allocated to care quality across the individual phases of care and overall for the entire episode of care. All three professional groups had moderate within-group inter-rater reliability, ranging from 0.46 (95% CI 0.34 to 0.59) to 0.52 (95% CI 0.41 to 0.62). The physician reviewers had the highest values. These results were replicated for the physician reviewers in the kappa analysis (Table 6a), although the kappa scores for the other two professional groups were considerably lower. The ICC reliability values were rather higher than the average found in a systematic by Lilford et al[5] in which, for implicit structured review, studies of case note review concerned with causality and process of care had kappa values below 0.4 (causality; kappa 0.39 (SD 0.19), process; kappa 0.35 (SD 0.19)).

The inter-rater reliability results of the study are also somewhat similar to those of Hofer and colleagues,[10] who also used Intra-Class Correlation Coefficients and found a reliability of 0.46 for a structured holistic review of diabetes and heart failure records by physician reviewers (though only 0.26 for COPD records). In comparison, a recent holistic assessment of patients dying in UK hospitals achieved a Kappa score of 0.39 on the key indicator of quality of medical care.[41] Nevertheless, our study found that there was still a considerable range of reliability scores between individual reviewer pairs, even between the doctor reviewers. This variation might be a reflection of either training or experience, or perhaps it represents other aspects of the causes of holistic review variation identified in the literature, such as bias or harshness in reviewing.[12][13][14][15][16]

Most of our reviewers had not undertaken a formal holistic review before. Training and support was provided but it was deliberately designed not to influence the effect of implicit judgement and the training may therefore not have been sufficient to reduce the element variation due to inexperience with the review methods. The large quality and safety review programmes in the United States (for example in the Colorado and Utah study[5] and the study by Daley and colleagues[37]) use senior physician reviewers who have been screened to identify and exclude those who may have particular biases. This approach might have had a positive impact on the inter-rater reliability results but was not an option open to this study because of cost and availability of senior staff.

Completeness of the clinical record is a significant feature in the success of review methods, since a review can only consider what is written in the record and reviewers have to depend on the abstracted details of the case. Not recording an event does not mean that the event did not occur. It may be, for instance, that a practitioner did not record the event because they considered it to be too trivial. However, non recording or very limited recording is a definite constraint on the effectiveness of records review as a means of assessing quality of care among groups of patients. On the other hand, direct observation of the quality of care, suggested as an alternative to records review,[9] would be very expensive if it were to be used as a standard procedure. Nevertheless, reviewers judged the quality of recording in the case notes as reasonable (over 4 on a score of 1-6) although some notable poor exceptions were commented on in the textual reviews.

The levels of inter-rater reliability for criterion based review ranged from moderate (0.61 for non-clinical audit staff) to quite high (clinical staff 0.74, doctors 0.88) and are similar to those found in the large UK national clinical audit programmes of stroke care[24][25] and continence care.[42] All three staff types performed equally well at capturing criterion based data from records, despite differences in their backgrounds, again confirming the findings of UK stroke care audit.[24][25] It is unsurprising that criterion based review has higher levels of inter-rater reliability than holistic review since the criteria are pre-determined, directly evidence based, have been subject to peer review and are explicit rather than implicit. Under these circumstances an ICC of less than 0.75 might be deemed unsatisfactory and any large study using explicit criteria might best train reviewers with this target in mind.

Criterion based review demonstrated that the reviewers from all three groups could identify relevant data where it existed (the effectiveness of reviewer scores were around a mean of 95%) and that in general sufficient case note data were available from which to assess quality of care using the review criteria. The quality of care scores for both criterion based review and from holistic review using overall scale scores were similarly rated by the three staff types at between 70% and 79% (Table 10), where 100% represents excellent care.

We found that in the hands of medically trained and of non-clinical audit reviewers, although not for the nurse and other clinical group, there was a significant level of agreement between the results of holistic and criterion based review, suggesting that the two methods are measuring related elements of quality. This may have important implications for the choice of review method when evaluating the quality of care from case notes. Criterion based review methods may give sufficient information on which to judge the overall quality of care, performing as a screening tool for large clinical audit studies, provided that the appropriate review criteria are chosen. On the other hand, a structured form of holistic review such as described here can, in the right hands, also give a reliable picture of the quality of care and pick up the nuances of quality variation that criterion based review is unable to provide.

*What additional contribution does holistic textual data make to the assessment of quality of care?*

Although there was some overlap between the results from the three professional groups, across the spectrum from descriptive statements to judgements some distinctive patterns also emerged (Tables 11 & 12, Boxes 5 & 6). It is not surprising that in general the reviewers without any clinical training tended to provide mainly descriptions of care process and only relatively rarely offered judgements on quality. The nursing group offered some judgements, especially about process or care issues, and in general were more concerned about the nursing process and the care plan than about the interventions offered to the patient. The doctors, on the other hand, were in general more concerned with assessing the quality and safety of the therapies and strategies for dealing with the acute illness rather than about the overall care plan.

Nurses tend to be close to the hospital medical care process because of frequent contacts with the patient so it might have been expected that the results from the nursing/clinical staff reviewers might have been close to those of the medical reviewers. But the limited agreement between the doctors and the nurses may also reflect different internal professional standards for assessing quality and safety of care when reviewing a record. Weingart and colleagues[6] conjectured that nurses and doctors reviewed in different ways, that nurses sought data on the routines of care while doctors looked for a wider picture and that, in general, neither group considered both dimensions. Analysis of the textual data tends to support the notion that the doctors and nurses in this study commented on different aspects of care when assessing quality. Judgements by the nurses tended to be implicit rather than explicit and they tended towards reviews of nursing process of care. Our doctors were the group who tended to mention outcomes or impact on future care, more so than the other groups, on the whole using explicit statements. Some were willing to justify why they thought care was good or unsatisfactory. Our results suggest that the hypothesis posited by Weingart et al[6] may have some validity – in reviewing quality of care nurses tend to concern themselves with care processes and pathways while doctors tend to be concerned with diagnosis and interventions. Each professional type may therefore identify nuances of care that the other does not.

Mean review times for both methods were quite similar for the doctor and nurse reviewers and the costs for the doctor group were consequently higher for each review. The cost differences between the nurse reviewers and the doctor reviewers were rather less for holistic review than for criterion based review. So the decision on which reviewer type to use for a review process will principally depend on which type of information is required from the review, though with an eye to the cost differentials. Because the cost of analysis of the textual data is high and requires specialist skills, and because there is some evidence from the relationship between criterion review and holistic scale score data, it is likely that any large scale study using holistic review would use scale scores to judge quality of care, rather than a full textual data analysis. On the other hand, for smaller scale, more detailed studies, analysis of textual commentary on quality of care provides a very rich data set on which to judge quality and safety.

Examination of the more than 100 reviews of cases judged to be unsatisfactory shows very few defined adverse events, the commonest of these being decisions to give the wrong drug.[43] Other, fewer, more serious events include missed or erroneous diagnoses. Much more commonly in the reviews there appear unsatisfactory aspects of care which may run as a thread throughout the hospital admission. What might be regarded individually as ‘minor shortcomings’ on their own become fused together to create an ‘event’. These shortcomings do not always translate across all of the phases of care. Sometimes, for example when care is handed over from one team to another, missed diagnoses or inappropriate or sub-optimal treatment (for example not following evidence based practice) is seen to be recovered by the receiving team and overall care is judged satisfactory. From our study we have been able to show that adverse events are at least as likely to be non-discrete constellations of sub-optimal components of care that, taken together, put the patient at risk. Not withstanding the additional costs, narrative descriptions of care through holistic analysis can considerably enhance understanding of healthcare quality and may be of particular value in locally based clinical audit.

*What is the relationship between mortality (outcomes) and quality of care scores for the study hospitals?*

In a systematic review of the relationship between quality and outcomes of care (risk-adjusted mortality) in hospitals or hospital units, Pitches and colleagues[38] found an uncertain relationship emerging from a heterogeneous group of 36 studies, including 51 quality and outcome relationships. About half of the studies demonstrated a link between better quality of care and the lower risk hospitals, about a quarter of studies showed a negative association and the remainder were equivocal.

Our study also found some mixed results (Table 15 and 16 and Figures 8 - 11). Across all staff types, for holistic review of cases of COPD, there was a trend for a positive relationship between higher overall quality of care scores and a lower mortality ranking, with one significant difference scored by the physician reviewers. For heart failure quality of care there was a trend towards higher scores in higher mortality hospitals, although the mean score differences were small. For the criterion based review there was a general trend towards higher quality scores for lower mortality hospitals, for both conditions it was only the results from the nurse/clinical reviewers for COPD, and only the physician reviewers for heart failure that showed a positive association with lower mortality.

It is likely that these associations are influenced by review method and staff type as well as by the method of risk adjustment and the actual quality of care provided and recorded. It could be argued that the differences between hospitals that might be found using criterion based review would be (and were) relatively limited because of the very structured nature of the criteria. So review criteria might be less useful in assessing differences between hospital units than the more broadly based holistic method. In contrast, triangulation of the holistic review intra-rater consistency, inter-rater reliability and qualitative analysis results suggests that the doctor reviewers on the whole produce information that does allow judgements about quality of care and that for both COPD and for heart failure there is a significant positive relationship between their overall holistic scale scores and mortality ranking of the hospitals– suggesting that for these two conditions better quality is found in hospitals with lower mortality.

Hofer and colleagues[10] suggested that inter-rater reliability results for chronic diseases such as heart failure and COPD were to some extent influenced by the evidence base, proposing that in their study the heart failure reviews had higher inter-rater reliabilities than COPD reviews because the evidence base for heart failure management was stronger than that for COPD. There is room for debate on this hypothesis but in any event we found that quality/mortality relationship was found more consistently across reviewers for COPD than for heart failure.

*Study limitations*

In a complex study such as this there are bound to be methodological limitations. There were only two tracer conditions used in the study whereas comparable studies in the United States have used five or more conditions.[10] Additionally, because of the nature of the research questions, only 38 reviewers in nine hospitals were involved and they came from a range of backgrounds, though with some similarity to the study by Weingart and colleagues.[6] Results from this study nevertheless show enough similarities with the results from Hofer et al[10] and Weingart et al[6] to suggest that the results are meaningful.

Assessment of the quality and safety of care using, among other methods, a six or ten point scale, remains unusual in the literature. Evaluation of the sensitivity of the scales was not possible prior to the main data collection and although the intra-rater reliability results suggested that there was a fair degree of internal consistency when these scales were used, further research on the use of similar scales would be of value.

There is a potential for bias in the study method in that the design, and especially the constraints of the ethics and research governance requirements, meant that reviewers evaluated the case notes in their own hospitals (rather than, for example, reviewing the case notes in another hospital where they would not have the possibility of reviewing cases in which they may have provided care). This is an acknowledged potential bias, although the range of holistic scores and the types of commentary that were recorded suggest that many of the reviewers were quite robust when determining the score for a case.

The reliability estimates reported in this study are likely to be larger than the true population reliability values because of the sampling method employed in the design. The variability of the results of the holistic reviews, including up to 15% of reviews scored 1 or 2, means that the ICC will tend to be higher than if the review results had been more homogeneous. Furthermore, the ICC ‘combines three features of the data (patient variability, reviewer variability and measurement variability) from which it is calculated….it does make comparisons of ICCs between different studies difficult to interpret’.[44] Nevertheless, a recent systematic review of the inter-rater reliability reported in case note review studies[11] was able to make kappa comparisons and this study had higher ICC results than the review kappa for both causality and process of care (causality; kappa 0.39 (SD 0.19), process; kappa 0.35 (SD 0.19)).

We have undertaken a relative simple statistical analysis to compare the reliability of the holistic and criterion based reviews by using kappa and ICC statistics. These statistical methods do not allow for other facets of measurement such as the disease or reviewer characteristics. The lack of balance in the design precludes a classic ANOVA based generalisability analysis to assess reliability. However, a multilevel modelling approach could be employed. We have not been able to undertake this but we acknowledge that such a model might produce smaller standard errors and give narrower confidence interval estimates for the reliability coefficients. However, we believe that this model is unlikely to change way the conclusions of the study in a substantive.

**Implications for reviewing quality of care**

To return to our practical question, what do our results tell us about which method of review would be best used for which purposes and by which professional groups? We have found that all three professional groups perform well when using criterion based review. If this type of review is chosen, perhaps for large scale clinical audit to inform service development, the decision on which professional groups should undertake reviews using this method might depend mainly on cost and availability of staff. The data on resource use show that the doctors are considerably more expensive at cost per review, because they have the highest salary levels, although their review times are similar to the nursing/clinical group. However, clinically trained reviewers may have a place when using structured review methods to identify variations in care such as adverse events, for clinical training might be of advantage by enhancing watchfulness. Furthermore, review of small numbers of cases has relatively little cost impact and for criterion based review of limited numbers of cases, for example in an investigation of quality of care, staff with nursing or medical training may add value to an evaluation.

The decision on who should undertake structured holistic review is more complex. It is a method that might best be used when more is required than just the sum of the results of collecting a set of review criteria. While all groups can use the method of holistic scale scoring, the overall results conceal wide ranges of agreement, sometimes close to random for phase-of-care results (Table 7). Particularly for the more technical phases of care such as investigations and admissions, and initial management, these results suggest that the three groups of staff are interpreting the recorded care differently when they each review the same record. This probably reflects their background knowledge of the clinical situation and of how the care is delivered. Even when considerable training in the review method has been provided, it is unrealistic to expect the non-clinical audit staff to fully appreciate the details of the medical care, let alone when that care has or has not deviated from best practice. Results suggest that nurse trained reviewers sometimes identify different problems from those found by physician reviewers. It is possible that extended training and selection of staff might reduce this difference, for instance by selection of specialist consultant nurses or of very experienced doctors. Selection of person skills according to task might provide the best outcome for the more difficult reviews.

Moreover, reviews of cases of serious unintended incidents or of poor outcome cases might benefit from structured reviews by pairs of reviewers, one with a nursing background and one with a medical background. Our results have shown that these two groups of reviewers offer different types of results, with nurses tending towards care process issues and doctors offering judgements on more technical interventions. If reviews were supported by effective training, including the enabling of staff to make explicit judgements on care, joint mixed professional reviewing, perhaps using more senior doctors and consultant level specialist nurses, might offer a wider range of insights than if case records were reviewed by two professionals from the same background. Whereas Weingart et al[6] wondered whether the differences in holistic review results from physicians and nurses reviewers could be problematic, the differences we found in our study could be put to a positive advantage.

Textual data provides much finer grained information than do scores or criterion based review, even when it is provided in short phrases and sentences. Full analysis of the textual data in a clinical setting, rather than in a research project, is likely to be costly and difficult to do when undertaking large scale audits or quality and safety reviews. However the increasing practice of undertaking smaller scale reviews, for example where there are a small series of cases with poor outcomes that require detailed review, are identifying a need for structured reviews that would benefit from a combination of data such as is provided by criterion based scores, holistic scoring and structured textual commentary.

Overall, the results of this study suggest that there may be significant gains to be made in clinical audit and evaluation through better understanding of the products of the different methods of review and of the value in training and selection of reviewers.

**What is the relationship between information on quality of care from case notes and hospital level outcomes of care?**

**Background**

The purpose of this second part of the study was to investigate how much of the variance in risk-adjusted outcome for two important clinical conditions could be attributed to differences in quality of clinical care in acute hospitals, as assessed through review of case notes.

Using case notes as the basis for assessing health care quality is known to be problematic for a number of reasons such as reliability,[10][12][45][6] bias[15] and consistency,[8] even though case notes are still almost universally used as a primary data source for this purpose. This continuing usage is partly due to the fact that the alternatives to using case note review as the basis for quality assessment – such as direct observation, use of video or actor-patients - may be even more time consuming and expensive and have their own methodological challenges. Approaches to reducing uncertainty in the measurement of quality of care have been concerned with improving reliability in implicit review, such as using structured review methods,[28] and providing evidence based review criteria for explicit review.[19][21]

In addition to measuring the process of care, the putative relationship between process and outcome is also an important consideration for healthcare systems and is the subject of scrutiny by public bodies.[46] Case note review has been used in a considerable number of studies to provide process of care data in order to explore the relationship with outcome of care, a relationship that appears complex. This was recently demonstrated by Pitches et al, [38] who undertook a systematic review of 36 studies, which included 51 ‘process-versus-risk-adjusted outcome relationships’, exploring the extent to which variations in risk adjusted mortality rates were associated with differences in quality of care. The authors found a positive correlation in only 51% of the relationships with no correlation in 31% and an unexpected correlation in a further 18%, in what was a very heterogeneous set of studies.

In a study of 87000 surgical operations in eight subspecialties in 44 hospitals with a range of risk adjusted mortality, Gibbs et al[7] used structured implicit review of case notes to populate the primary outcome measure of a five point scale that assessed overall quality of care. Overall, the authors found that quality of care ratings were not significantly different between hospitals with higher than, and lower than, expected mortality.

National quality of care audits in the UK, such as the National Stroke Audit,[47] have used explicit review criteria to assess quality of care. Using a 60 item set of criteria derived from the UK stroke audit in a longitudinal case note review study of stroke care in New Zealand, McNaughton[48] and colleagues found only weak relationships between process and outcome variables across four hospitals and the hospital with the best process scores had the worst case mix-adjusted outcomes. A study of 20 UK maternity units found substantial change in practise over eight years but few associations between proxy outcomes and other explanatory variables.[49]

Earlier in the study we considered the benefits and limitations of using case notes as the basis for reviewing quality of care and examined the reliability and value of two review methods – holistic (implicit) review and criterion based (explicit) review - in the hands of different types of health professional. We showed that the two methods provide different types of quality assessment with reasonable to good levels of reliability.

In this next stage of the study, hospital based process of care is assessed using mixed case note review methods, with implicit review structured by phases of care, including both scale scores and structured textual data, together with explicit, criterion based, scores - and outcome is judged using a set of direct and indirect measures derived from national data sets.

***Study aim and research questions***

***Aim:*** To investigate how much of the variance in outcome for two important clinical conditions (adjusted for measurable differences in risk) can be attributed to known differences in quality of clinical care in acute hospitals.

***Research questions***

What is the relationship between the quality of care for individual conditions in hospitals and overall quality of care and quality markers across hospital institutions?

What is the relationship between casemix or risk adjusted outcome and quality of care as measured by case record review?

Is high quality care associated with in high quality outcomes in risk adjusted cases?

Is there a correlation in clinical quality between the management of different conditions in the same hospital, as measured by case note review?

**Methods**

**Choice of settings, review methods and staff**

***Selection of study hospitals***

In our original proposal we indicated that a total of 24 acute hospitals in England would contribute data to this process/outcome study, including the data from with the case note reviews in the initial eight (nine) hospitals in the reliability study. During the analysis of the first part of the study, it became clear that the method of data collection being selected for the outcomes study - using a single physician reviewer per condition per hospital and a compound method of reviewing - meant that the data from the reliability study could not contribute fully to the objectives of this outcome study. Thus a total of 24 hospitals were separately recruited for this quality and outcomes study.

Selection of the hospitals used the same processes in the phase one reliability study. First, Hospital Episode Statistics[33] on 28 day mortality data for COPD and Heart Failure were accessed through the East Midlands Public Health Observatory. Hospitals were excluded from the selection process if they reported less than 200 in-patient cases per year for either condition, effectively excluding smaller or specialist acute hospitals. There were 136 hospitals in the final data set.

Second, twenty eight day mortality data for the two study conditions for each hospital was combined using simple averaging, to create an average 28 day mortality ratio for each hospital. Third, these were then ranked from the lowest mortality to the highest and the data was split into four quartiles. And finally, from this ranking, hospitals were randomly selected in each of the lowest and the highest mortality quartiles.

The 20 hospitals selected for the reliability study were excluded from the selection process. Thirty hospitals were then randomly selected from the remaining 116 in the data set, 15 from the lower mortality quartile and 15 from the higher mortality quartile. Six additional hospitals over the proposed 24 were selected to take account of the likely drop-out rate during the recruitment and fieldwork phases of the study.

At the commencement of data collection there were 25 hospitals included in the study. However, only 20 hospitals in total, 10 each in the upper and lower mortality groups, were able to collect data on both COPD and heart failure. Thus only 20 hospitals were included in the overall analysis.

Figures I & 2 demonstrate the mortality differences between the two groups of hospital included in the final analysis, each mortality group (upper and lower) containing 10 hospitals.

|  |  |
| --- | --- |
| **Figure 1: Box plot of COPD/Heart Failure mortality showing differences between higher and lower mortality groups** | **Figure 2: Bar chart showing individual hospital level mortality rate differences** |
|  | **Hospitals with lower mortality, n=10**  **Hospitals with higher mortality, n=10** |

***Selection of review conditions***

This quality and outcomes study used the same two conditions as in the phase one study – namely, admissions for acute exacerbation of COPD and of heart failure – in each study hospital. These working definitions for data collection were (Box 1):

**Box 1**

Definition of an exacerbation of COPD:

“An exacerbation is a sustained worsening of the patient’s symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, cough, increased sputum production and change in sputum colour.” [31]

Definition of an exacerbation of Heart Failure: An exacerbation of heart failure is a sustained worsening of the patient's symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, tiredness and swelling of the feet and/or ankles." [22]

*N****umber of case notes for review and of reviews***

Analysis of the data from the first part of the study suggested that 40 case notes would be an appropriate number to review at each hospital, for each condition, in a study that was to seek associations between recorded quality of care and hospital level outcomes. Furthermore, review of 40 case notes has been shown in the UK Stroke and COPD national audits[47][34] to be practical and to yield useful data. In contrast to the earlier analysis, it was not intended to undertake intra-rater and inter-rater reliability analyses, so it was not necessary for each set of records to be reviewed more than once.

In our proposal for phase two of the study we said that in order to test the reliability of the review process, 10 sets of case notes per condition per hospital would be double reviewed (rather than double reviewing all case notes). However this target of 10 double reviews proved not possible, partly because of the costs involved in undertaking the double reviews and partly because of inability to recruit second reviewers of the same staff type, beyond the 50 reviewers already required.

***Choice and recruitment of reviewers***

In our initial research proposal we said that only one review method would be used at this stage of the project, the choice being made following the analysis of the data in the reliability study. Analysis of the reliability study showed that while there were similarities between the criterion based reviewing skills of the doctor and nurse, there were differences in the type of holistic data captured by the two professional groups and which formed the basis of the judgements about quality of care scores. Medically trained reviewers tended to judge the quality of interventions while nurses tended to review from a care pathway perspective. Since the study reported here sought associations between process of care and outcomes of care, it was judged that medically trained reviewers in higher specialty training were likely to provide the most appropriate data.

Following the random selection of candidate hospitals, in each hospital a senior physician working in cardiology (heart failure care), and a senior physician working in COPD care, was asked to recruit a physician reviewer in mid stage specialty training in their relevant specialty.

***Selection and refinement of review methods***

One of the subsidiary purposes of the earlier reliability study was to determine which of two review methods were most appropriate for the quality of care/outcomes study - criterion based review or holistic review. We had found that the two methods were complementary, although different, and that therefore both methods might have a value in the review of care for the outcomes study. Review of care using evidence based criterion methods would provide information on the extent to which care, overall, fitted with external standards. Structured holistic review with quality of care scores, together with short explanatory comments where reviewers thought it appropriate, would provide information about the nuances of care that could not be identified through the use of pre-constructed review criteria. In this second part of the study it was not necessary to separate out the reviewing stages (holistic review followed by criterion based review), since the results of the two methods were not being compared one with another, as in the phase one study. It was therefore possible to merge the holistic and criterion based reviews into one process so that the case notes were only reviewed once and a single data collection form could be used.

Using a form of structured implicit review, the reported quality of care provided to each patient was scored for each of three phases of care (admission/investigations, initial management and pre-discharge care) and for care overall.

In the earlier part of the study holistic review a six point Likert scale was used by reviewers to rate the quality of care in each of the three phases of care, together with a 10 point Likert scale which was used to rate the overall quality of care. Although the 10 point scale was more finely grained, comparison in the phase one study between the six point phase scores and the 10 point scale added to the complexity of the analysis without providing obvious benefits to the structured review process. A six point scale was therefore used for all phases and overall care review in this study, with anchors of 1 = unsatisfactory and 6 = very best care. The descriptors for the six points are shown in Box 2.

**Box 2**

1. Care fell short of current best practice in one or more significant areas resulting in the potential for, or actual, adverse impact on the patient
2. Care fell short of current best practice in more than one significant area, but is not considered to have the potential for adverse impact on the patient
3. Care fell short of current best practice in only one significant area but, is not considered to have the potential for adverse impact on the patient
4. This was satisfactory care, only falling short of current best practice in more than two minor areas
5. This was good care, which only fell short of current best practice in one or two minor areas
6. This was excellent care and met current best practice.

Two prompting questions had earlier been used to seek textual comments on the quality of care but reviewers tended only to respond to one or other of the questions. Thus in this study, only one question was asked of reviewers when they provided textual comment on the quality of care. Box 3 shows the format of the questions used for phases of care and care overall.

**Box 3**

**Investigations/examination (for example)**

We are interested in comments about the quality of care the patient received and whether it was in accordance with current best practice (for example, your professional standards). You may also wish to comment from your own professional viewpoint. If there is any other information that you think is important or relevant that you wish to comment on then please do so.

Please comment on the care received by the patient during this phase.

Please rate the care received by the patient during this phase.

Please tick only one box

Unsatisfactory □ □ □ □ □ □ Very best care

The quality of care provided was also measured through the presence of a condition specific set of review criteria for each of the two study conditions. A number of changes to the earlier data collection methods were required to make the criteria and data capture tools fit for purpose for the outcomes study. During the first phase study it became apparent that a limited number of the review criteria had far higher levels of missing data than other criteria. We presumed that this was because the data was routinely unavailable, or more difficult to find, or that the criteria were more difficult to understand. Whatever the reason, these criteria added nothing to the review process and were removed. Additionally, because of the potential impact of poor quality recording on the ability to assess quality of care, reviewers were asked to rank the quality of the information held in the clinical records on a Likert scale from 1 (poor) to 6 (excellent).

Reviewers were unaware of their hospital mortality ranking – whether they were in the higher or lower mortality quartile – and they evaluated the records within their own hospital, as would happen with local clinical audit. No patient-identifiable data was returned to the study team or used in the analysis.

***Reviewer training***

Reviewers were trained to use a combination of two forms of case note review – criterion based (explicit) review and holistic (implicit) review - and also to provide written critical commentary on quality of care received by the patient, including on adverse events. The majority of reviewers were trained at a seminar in which the review methods were presented, using a set of anonymised case notes as the basis for small group and larger directed discussions. Topics discussed included techniques to find relevant information and group discussion to identify good and less satisfactory performance from the case notes.

Some reviewers were unable to attend the main training sessions and they received one-to-one or small group training from study research staff.

**Selection of outcome data**

An *a priori* choice of outcome measures relevant to the study was made by the multi-professional study team. Measures were selected from a range of sources, relating to:

* Clinical practice measures, such as 28 day mortality rates for COPD and heart failure (from Hospital Episode Statistics[33]) and patients with myocardial infarction receiving thrombolysis within an hour.[46]
* Hospital level proxy clinical outcome data such as Hospital Standardised Mortality Ratio (HSMR).[50]
* Proxy measures of patient safety and safety climate (for example from the NHS Staff Survey 2006[51] and the National Patient Safety Agency (NPSA),[52] used in an evaluation of incident reporting levels to the Agency. These data items were selected from a larger set as a result of previous empirical research.[53]
* External review data such the ability of the hospital to meet national targets for quality, collected by the Healthcare Commission for England.[46]

The final set of outcome variables is shown in Box 4.

**Box 4: Direct and proxy outcome hospital level variables included in the analysis**

|  |
| --- |
| **Variable** |
| Percentage of COPD or HF patients who die in hospital within 28 days |
| Hospital SMR (3 year mortality) (Dr. Foster) |
| Hospital SMR (1 year mortality) (Dr. Foster) |
| Numbers of incidents reported by the hospital to the National Patient Safety Agency per 100 bed days |
| SMR for deaths in low mortality Healthcare Related Group’s (HRGs) |
| COPD or HF finished consultant NHS episodes (Hospital Episode Statistics (HES)) |
| COPD or HF bed days (HES) |
| COPD or HF mean length of stay (HES) |
| COPD or HF mean age (HES) |
| Healthcare Commission for England Star rating (0 worst to 3 best) |
| Use of resources (Healthcare Commission) |
| Patient’s experience (Healthcare Commission) |
| Quality of services (Healthcare Commission) |
| Percentage of patient’s with acute myocardial infarction receiving thrombolysis (Healthcare Commission) |
| Extent to which hospital meets existing national targets (Healthcare Commission) |
| Extent to which hospital meets new national targets (Healthcare Commission) |
| NHS staff survey Q25a: Have seen errors in the past month (% yes) |
| NHS staff survey Q27b: Encouraged to report errors (mean) |
| NHS staff survey Q27e: The hospital takes action to ensure incident does not happen again (mean) |
| NHS staff survey Q24a: I know how to report (% yes) |
| NHS staff survey Q24b: I know the system for reporting (% yes) |
| NHS staff survey Q24b: I know the system for reporting (% no) |
| NHS staff survey Q24b: I know the system for reporting (% don’t know) |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) |
| NHS staff survey Q22f: Happy with standard of care provided (mean) |

***Analytic approach***

Quantitative quality of care data from holistic scale scores and criterion scores were examined using summary statistics.

The outcomes data set for each hospital was constructed from the data available for each hospital on the data sets identified in Box 4 (above). For the ranked mortality analyses, data were combined for the 10 hospitals in each of the higher or lower mortality groups.

Correlation of the relationship between the quality of care (process) data and outcome data was undertaken using Pearson’s correlation coefficient for continuous data together with linear regression analysis. Multiple regression analysis was used where appropriate. Spearman’s correlation coefficient was used for ordered categorical data (such as quality of service ratings). Levels of strength of correlation are shown in Box 5.

**Box 5**

|  |  |
| --- | --- |
| |r| ≥ 0.8 | Very strong relationship |
| 0.6 ≤ |r| < 0.8 | Strong relationship |
| 0.4 ≤ |r| < 0.6 | Moderate relationship |
| 0.2 ≤ |r| < 0.4 | Weak relationship |
| |r| < 0.2 | Very weak relationship |

All statistical tests were two-sided and a significance level of p ≤ 0.05 was used for all analysis except multiple regression analysis, where a level of p ≤ 0.1 was used.

Qualitative data from the textual commentary was examined by four reviewers, three of whom analysed data from 17 reviews each (of 40) to ensure some overlap, while a fourth reviewer reviewed a further sample of five from the total. Consistency among reviewers was checked after all had assessed output from two reviews, followed by discussion of the results. Further consistency check was undertaken between all four reviewers prior to completion of the analysis.

The results of each qualitative assessment of a reviewer’s output, in which analysts identified cases being described as excellent, good, satisfactory, unsatisfactory or very unsatisfactory, were then checked against the overall scale scores to judge whether a reviewer’s quantitative scores were reflected in their qualitative description of the quality of care provided for a case.

**Results**

Within the 20 study hospitals, reviews of case notes were undertaken of 873 people with COPD and 692 people with heart failure (1565 in total). The numbers of reviews undertaken by each reviewer for COPD ranged from 8 - 40 (median 40, mean 35) and for heart failure from 10 - 49 (median 40, mean 35).

Reviewers were asked to assess the quality of recording in the case notes because of the potential impact of poor quality recording on the ability to assess quality of care. On a scale of 1-6 the quality of case notes was scored satisfactory or better in 85% of the reviews

**Relationship between outcome variables and higher mortality and lower mortality groups of hospitals**

There were 10 hospitals in each mortality group. Table 1 presents the mean scores for continuous outcome variables available from national data sets (excluding NHS staff survey outcomes) for the higher and lower mortality groups of hospitals. There was a significantly higher percentage of both COPD and heart failure patients who died within 28 days in the higher mortality group of hospitals compared with the lower mortality group of hospitals (as would be expected from the hospital selection criteria) and the 1 year mortality data available from the ‘Dr. Foster’ analysis[50] was significantly higher for higher mortality group of hospitals.

Additionally, the mean age for heart failure patients was significantly higher for higher mortality group of hospitals compared to the lower mortality group of hospitals. The differences in mortality rates between the two groups of hospitals were as expected from the selection criteria for those hospitals, so that the results serve to confirm the expected differences (see also Figures 1 & 2). The one year mortality data from the Dr Foster source also support the expected differences in mortality and the higher mean age of people with heart failure in the higher mortality group of hospitals may be a partial explanation for those higher mortality rates. No other significant associations were found.

**Table 1: Mean scores for continuous outcome variable by hospital mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Higher mortality  (SD) | Lower mortality  (SD) | t-test  (p-value) |
| HSMR from Dr. Foster (3 year mortality) | 107.1 (12.2) | 99.2 (8.6) | 1.67 (0.112) |
| HSMR from Dr. Foster (1 year mortality) | 106.3 (12.6) | 95.6 (8.9) | 2.20 (0.041) |
| Incidents to NPSA per 100 bed days | 0.74 (0.79) | 0.70 (0.72) | 0.13 (0.896) |
| SMR for deaths in low mortality HRG’s | 92.7 (41.2) | 113.9 (38.0) | -1.20 (0.247) |
| COPD finished consultant episodes | 1164 (1012) | 1427 (815) | -0.63 (0.540) |
| COPD bed days | 7170 (5117) | 7268 (3744) | -0.05 (0.962) |
| COPD mean length of stay | 9.4 (1.8) | 8.7 (2.5) | 0.68 (0.507) |
| COPD mean age | 69.2 (1.3) | 69.1 (2.6) | 0.12 (0.905) |
| HF finished consultant episodes | 636 (525) | 765 (525) | -0.53 (0.601) |
| HF bed days | 5298 (3425) | 5680 (4025) | -0.22 (0.828) |
| HF mean length of stay | 12.8 (1.0) | 12.5 (1.1) | 0.59 (0.561) |
| HF mean age | 78.1 (1.5) | 75.9 (2.3) | 2.48 (0.024) |
| Patient’s experience | 68.2 (3.7) | 67.2 (4.5) | 0.58 (0.567) |
| Percentage of patient’s with acute MI receiving thrombolysis | 56.3 (18.9) | 56.6 (21.3) | -0.04 (0.967) |

Correlations at the < 0.05 significance level highlighted

***Relationship between mortality groups and quality of care***

Table 2 shows that most reviewers classify care in the good range (about 41% for COPD, and 46% for Heart Failure, (score 5)) with around 19% - 21% in the ‘fell short of best practice’ range (score 1 to 3) and the lowest scores (1, unsatisfactory) under less than 5% for both conditions across the 20 hospitals.

The relationship between the holistic data (including mean scale scores for overall care, all of the phase of care scores) and the mean total criterion score, correlated with the hospital groups ranked by mortality, was analysed using the two independent samples t-test (Tables 3 & 4). Mean difference is the mean score for each case note review in the higher mortality group of hospitals minus the mean score for the lower mortality group of hospitals. A negative mean difference in the table indicates that the lower mortality group mean score is higher than the mean score for the higher mortality group.

For both conditions the findings are therefore consistent. Across all phases of care and overall care using holistic review, and total criterion score, there was a trend towards higher mean scores for the lower mortality hospital groups. However, when account was taken of clustering, there were no statistically significant differences in scores between the two groups of hospitals.

**Table 2: Distribution of overall holistic quality of care scores: total number (and percentage) of reviews in each category**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fell short of good practise | | | Satisfactory | Good or better | |  |
| Scale scores | 1 (unsatisfactory) | 2 | 3 | 4 | 5 (good) | 6 (very best care) |  |
| Numbers of reviews |  |  |  |  |  |  | Total numbers of reviews |
| COPD  (%) | 38  (4.4) | 56  (6.4) | 87  (10.0) | 166  (19.1) | 362  (41.3) | 164  (18.8) | 873  (100%) |
|  |  |  |  |  |  |  |  |
| Heart failure  (%) | 26  (3.8) | 44  (6.3) | 60  (8.7) | 154  (22.2) | 318  (45.9) | 91  (13.1) | 692  (100%) |

**Table 3: Relationship between the COPD mean holistic scale scores (1-6) scale, total criterion scores and hospital mortality group**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Lower mortality group scores | | | Higher mortality group scores | | | Mean  Diff | 95% CI | | P-value |
|  | N | Mean | SD | N | Mean | SD |  | Lower | Upper |
| Total Criterion score | 438 | 10.3 | 1.7 | 447 | 10.1 | 1.8 | -0.2 | -0.9 | 0.6 | 0.637 |
| Holistic Admission phase | 434 | 4.7 | 1.2 | 441 | 4.5 | 1.4 | -0.2 | -0.6 | 0.2 | 0.321 |
| Holistic Initial Management Phase | 435 | 4.8 | 1.2 | 438 | 4.7 | 1.2 | -0.1 | -0.4 | 0.2 | 0.649 |
| Holistic Pre-discharge phase | 423 | 4.8 | 1.1 | 423 | 4.7 | 1.2 | -0.1 | -0.4 | 0.2 | 0.542 |
| Holistic Overall quality of care rating | 428 | 4.5 | 1.3 | 444 | 4.3 | 1.3 | -0.2 | -0.6 | 0.3 | 0.419 |

Mean difference is the mean score for case notes of patients in lower mortality group hospitals minus mean score

for the higher mortality group.

The P-values and confidence intervals were adjusted for clustering by reviewer.

**Table 4: Relationship between the HF mean holistic scale scores (1-6) scale, total criterion scores and hospital mortality group**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Lower mortality group scores | | | Higher mortality group scores | | | Mean  Diff | 95% CI | | P-value |
|  | N | Mean | SD | N | Mean | SD | Lower | Upper |
| Total Criterion score | 393 | 8.0 | 1.4 | 309 | 7.8 | 1.5 | -0.2 | -0.7 | 0.2 | 0.312 |
| Holistic Admission phase | 393 | 4.7 | 0.9 | 309 | 4.6 | 1.2 | -0.1 | -0.6 | 0.3 | 0.501 |
| Holistic Initial Management Phase | 382 | 4.6 | 1.1 | 304 | 4.5 | 1.3 | -0.1 | -0.6 | 0.3 | 0.541 |
| Holistic Pre-discharge phase | 384 | 4.7 | 1.0 | 300 | 4.3 | 1.4 | -0.4 | -1.0 | 0.1 | 0.122 |
| Holistic Overall quality of care rating | 385 | 4.6 | 1.1 | 308 | 4.2 | 1.4 | -0.4 | -0.9 | 0.1 | 0.115 |

Mean difference is the mean score for case notes of patients in lower mortality group hospitals minus mean score

for the higher mortality group.

The P-values and confidence intervals were adjusted for clustering by reviewer.

The similarities between the quality of care scores in the upper and lower hospitals are further demonstrated when the relationship between mean overall and mean phase holistic scale scores and ranked mortality is explored. Tables 5 - 8 show the relationships between holistic scores (ranging from 1 = unsatisfactory to 6 = very best care) and ranked mortality. No significant differences in scores were found between the upper and lower mortality groups of hospitals.

**Table 5: COPD - Mean holistic overall score by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality hospital group (N = 10) | 4.3 (0.62) | 4.4 | 2.9 to 5.0 |
| Lower mortality hospital group (N = 10) | 4.5 (0.56) | 4.5 | 3.7 to 5.5 |

There were no statistically significant differences in mean scores per mortality group (t = -0.83, p = 0.418).

**Table 6: Heart Failure -** **Mean holistic overall score by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality hospital group (N = 10) | 4.2 (0.58) | 4.1 | 3.1 to 5.1 |
| Lower mortality hospital group (N = 10) | 4.6 (0.38) | 4.6 | 3.9 to 5.1 |

There were no statistically significant differences in mean scores per mortality group (t = -1.75, p = 0.097).

**Table 7: COPD - Mean holistic phase score by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality hospital group (N = 10) | 13.4 (1.65) | 13.6 | 10.2 to 15.2 |
| Lower mortality hospital group (N = 10) | 13.9 (1.30) | 14.0 | 11.6 to 16.1 |

There were no statistically significant differences in mean scores per mortality group (t = -0.77, p = 0.453).

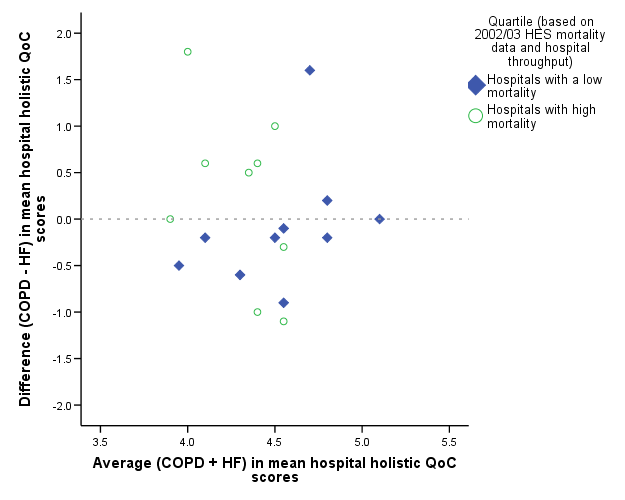
**Table 8: Heart Failure - Mean holistic phase score by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality hospital group (N = 10) | 13.0 (1.83) | 13.4 | 10.3 to 15.3 |
| Lower mortality hospital group (N = 10) | 13.8 (1.44) | 13.8 | 11.5 to 15.8 |

There were no statistically significant differences in mean scores per mortality group (t = -1.10, p = 0.288).

The relationship between of the mean holistic overall quality of care scores for COPD and heart failure in each hospital was also explored using a Bland-Altman plot (Figure 3) to investigate whether the two scores tended to be in the same direction in each hospital. That is, if the hospital was in the lower mortality group, which on average had somewhat higher group mean scores for both COPD and heart failure, was this pattern replicated within each hospital. No clear pattern emerged between the two mean scores for each hospital, grouped for overall mortality, so that it is not possible to see a trend towards higher scores for both conditions in each of the lower mortality hospitals, nor for the lower scores for each condition in each of the hospitals in the higher mortality group.

**Figure 3: Relationship between mean quality of care scores for COPD and heart failure within each hospital, grouped by mortality level**



Note;- one hospital in the higher mortality hospital group was an outlier in all of the analyses and was therefore excluded from the Bland-Altman plot analysis.

**Correlations between holistic scale scores and hospital level outcome variables**

Pearson’s correlation coefficient was used of analysis for continuous data. Spearman’s correlation coefficient was used for analysis of ordered categorical data (for example, Healthcare Commission star rating). Levels of strength of correlation are assumed as in Box 5 above (repeated here to assist the reader).

|  |  |
| --- | --- |
| |r| ≥ 0.8 | Very strong relationship |
| 0.6 ≤ |r| < 0.8 | Strong relationship |
| 0.4 ≤ |r| < 0.6 | Moderate relationship |
| 0.2 ≤ |r| < 0.4 | Weak relationship |
| |r| < 0.2 | Very weak relationship |

Rather than using statistical significance to identify important results, correlations were identified as important if the correlation coefficient was greater than 0.4; that is at least a moderate relationship existed between the two variables under investigation. Very weak and weak relationships (r≤ 0.4) were regarded as having no correlation.

There were only a limited number of moderate to very strong correlations between the quality of care scores for the upper and lower mortality hospital groups and the outcome and risk adjustment variables selected for the study. A positive correlation indicates that as criterion or overall holistic score increases the variable levels also increase. Thus, for example, a positive correlation for NHS staff survey question 22e (care of patient/service user is top priority) demonstrates an increase in COPD or heart failure criterion score as there is an increase in the mean score for staff who agree that care of patient/service user is top priority in the hospital.

An example of a negative correlation would be an increase in COPD or heart failure score which correlates with a decrease in the number of incidents reported to the National Patient Safety Agency per 100 bed days (increased reporting is thought to indicate a positive safety culture[53]).

Tables 9 - 12 present the moderate to very strong correlations only. The results of all of the holistic data correlations, including those without statistically significant results, are presented in Appendices 7 to 10. Pearson correlation coefficients are used unless otherwise indicated in the table. Each table is accompanied by scatter plots fitted with regression lines to demonstrate the strength of the correlation.

*COPD mean holistic overall score (1 unsatisfactory to 6 excellent)*

Table 9 shows that there were only two moderate negative correlations. Thus as the COPD scale score increased (an indication of better quality of care), the frequency decreased at which NHS staff said they were encouraged to report errors (Figure 3), and also when staff said that care of the patient/service user was a high priority (Figure 4). These correlations are in an unexpected direction since better quality of care might have been expected to relate to both a positive safety culture and to high priority for patient care.

**Table 9: Correlation between COPD holistic mean overall score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| NHS staff survey Q27b: Encouraged to report errors (mean) | -0.490 | 0.028 | Moderate |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | -0.503 | 0.024 | Moderate |

|  |  |
| --- | --- |
| **Figure 3.**  **Scatter plot (with regression line) for COPD score by NHS staff survey 27b: encouraged to report errors (mean score)** | **Figure 4.**  **Scatter plot (with regression line) for COPD score by NHS staff survey 22e: care of patient/service user is top priority (mean score)** |
|  |  |

*Heart failure mean holistic overall score (1 unsatisfactory to 6 excellent)*

A total of four variables were significantly correlated with heart failure mean score: quality of services; existing national targets; new national targets and NHS staff survey question 24a (know how to report concerns) (Table 10).

There is a strong positive correlation between quality of services (fair, good or excellent) and heart failure mean score: those trusts with a fair score had lower mean heart failure scores than trusts with a good or excellent score (Figure 5). Existing national targets (partially met, almost met, fully met) were also strongly positively correlated with heart failure mean scores where an increase in mean scores indicated an increased likelihood of trusts meeting existing national targets (Figure 6). There was a moderate positive correlation between new national targets (weak, fair, good, excellent) and heart failure - mean heart failure scores increased with increased levels of excellence (Figure 7).

Finally, there was a moderate positive correlation between NHS staff survey 24a and heart failure mean score:- the higher the percentage of responders who knew how to report incidents the higher was the mean heart failure score (Figure 8).

**Table 10: Correlation between heart failure holistic mean overall score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Quality of services | 0.651\* | 0.002 | Strong |
| Existing national targets | 0.765\* | < 0.001 | Strong |
| New national targets | 0.453\* | 0.045 | Moderate |
| NHS staff survey Q24a: Know how to report concerns (% yes) | 0.509 | 0.022 | Moderate |

\* Spearman’s rank correlation used

|  |  |
| --- | --- |
| **Figure 5.**  **Scatter plot (with regression line) for heart failure score by quality of service** | **Figure 6.**  **Scatter plot (with regression line) for heart failure score by existing national targets** |
|  |  |
| **Figure 7.**  **Scatter plot (with regression line) for heart failure score by new national targets** | **Figure 8.**  **Scatter plot (with regression line) for heart failure score by NHS staff survey Q24a: Know how to report (percent yes)** |
|  |  |

*COPD mean holistic phase score*

Only NHS staff survey question 22e (care of patient/service user is top priority) was significantly correlated with COPD mean holistic score (Table 11). As with the overall score analysis above (Table 9), the correlation between the two variables was negative, and as COPD mean holistic score increases the mean score for care of patient/service users being a top priority decreased – a lower score indicates more disagreement with the statement (Figure 9).

**Table 11: Correlation between COPD holistic mean phase score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | -0.454 | 0.044 | Moderate |

|  |
| --- |
| **Figure 9.**  **Scatter plot (with regression line) for COPD holistic score by NHS staff survey Q22e: care of patients/service users is top priority (mean score)** |
|  |

*Heart failure mean holistic phase score*

Seven variables were significantly correlated with heart failure mean holistic score:-

heart failure mean age (HES data), quality of services, percentage of patients with acute MI receiving thrombolysis, existing national targets, staff survey 27e (trust takes action to ensure never happens again), staff survey 24a (know how to report incidents) and staff survey 22e (care of patient/service user is top priority).

Mean age is negatively correlated with score:- as mean age increases mean holistic score decreases (Figure 10). This may be explained by older people having more complex problems, although reviewers in the phase one study showed that quality of care could be scored highly even when a patient did not survive.

Quality of service (fair, good, excellent) is positively correlated with mean score, thus as heart failure mean phase score increases the more likely a trust is to be of good or excellent status (Figure 11). There is a moderate positive correlation between the percentage of patients with acute MI receiving thrombolysis and heart failure mean holistic phase score (an important association in heart failure care) (Figure 12). A strong positive correlation exists between existing national targets (partly met, almost met, fully met) and heart failure mean holistic phase score, mean scores increase as the level of meeting targets increases (Figure 13).

Additionally, there were moderate positive correlations between mean score for questions 27e and 22e on the NHS staff survey and heart failure mean holistic phase scores the higher the heart failure score the stronger the respondent agreed with the statement that the trust takes action to ensure it does not happen again or care of patient/service user is top priority (Figure 14). Finally, there was a moderate positive correlation between staff survey 24a and heart failure mean holistic phase score, where the higher the percentage of responders agreeing (% yes) they know how to report incidents results the higher the heart failure mean holistic score (Figure 15).

**Table 12: Correlation between heart failure holistic mean phase score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| heart failure mean age | -0.552 | 0.014 | Moderate |
| Quality of services | 0.486\* | 0.030 | Moderate |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.463 | 0.046 | Moderate |
| Meets existing national targets | 0.691\* | 0.001 | Strong |
| NHS staff survey Q27e: hospital takes action to ensure does not happen again (mean) | 0.470 | 0.037 | Moderate |
| NHS staff survey Q24a: Know how to report (% yes) | 0.546 | 0.013 | Moderate |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | 0.470 | 0.037 | Moderate |

Spearman’s rank correlation used

|  |  |
| --- | --- |
| **Figure 10.Scatter plot (with regression line) for heart failure holistic score by mean patient age** | **Figure 11. Scatter plot (with regression line) for heart failure holistic score by quality of services** |
|  |  |
| **Figure 12.**  **Scatter plot (with regression line) for heart failure holistic score by percentage of patients with acute MI receiving thrombolysis** | **Figure 13.**  **Scatter plot (with regression line) for heart failure holistic score by existing National targets** |
|  |  |
| **Figure 14.**  **Scatter plot (with regression line) for heart failure holistic score by NHS staff survey 24a: Know how to report (percent yes)** | **Figure 15.**  **Scatter plot (with regression line) for heart failure holistic score by NHS staff survey Q22e: care pf patients/service users is top priority (mean score)** |

**Correlations between criterion based scores and hospital level outcome variables**

Tables 13 & 14 show the relationships between criterion scores and mortality grouping. There were no statistically significant differences in mean scores per mortality group. These findings are similar to the holistic quality of care data (Tables 5 – 8).

**Table 13: COPD mean criterion based score (out of 13) by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality group hospitals (N = 10) | 10.1 (1.10) | 10.4 | 7.9 to 11.5 |
| Lower mortality group hospitals (N = 10) | 10.2 (0.73) | 10.7 | 8.4 to 11.1 |

There were no statistically significant differences in mean scores per mortality group (t = -0.21, p = 0.840).

**Table 14: Heart failure mean criterion based score (out of 11) by mortality group**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (SD) | Median | Range |
| Higher mortality group hospitals (N = 10) | 7.8 (0.41) | 7.9 | 7.0 to 8.4 |
| Lower mortality group hospitals (N = 10) | 8.0 (0.50) | 7.9 | 7.5 to 9.3 |

There were no statistically significant differences in mean scores per mortality group (t = -0.83, p = 0.417).

***Correlations between criterion based scores and outcome variables***

Correlations at the moderate to very strong level are shown in tables 15 & 16, with accompanying scatter plots. The results of all correlations are presented in Appendices 11 and 12.

*COPD mean criterion score (out of 13)*

Two variables were significantly correlated with COPD mean criteria score: incidents reported to the NPSA per 100 bed days and NHS staff survey question 22f (happy with standard of care provided) (Table 15). A moderate negative correlation was observed with NPSA bed days, as the number of incidents per 100 bed days decreased the mean criteria score increased (Figure 16). Again this is an unexpected direction for the correlation, since increased reporting is a positive safety culture marker. A moderate positive correlation was observed with NHS staff survey 22f (mean score), where the higher the score the happier the staff with their standard of care, thus as the mean response to 22f increased, mean COPD score increased (Figure 17).

**Table 15: Correlations between COPD mean criterion score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Incidents reported to NPSA per 100 bed days | -0.489 | 0.029 | Moderate |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | 0.475 | 0.034 | Moderate |

|  |  |
| --- | --- |
| **Figure 16.**  **Scatter plot (with regression line) for COPD criterion score by incidents reported to NPSA per 100 bed days** | **Figure 17.**  **Scatter plot (with regression line) for COPD algorithmic score by NHS staff survey Q22f: happy with standard of care provided (mean score)** |
|  |  |

*Heart failure mean criteria score (out of 11)*

Two variables were significantly correlated with heart failure mean criteria score: incidents reported to the NPSA per 100 bed days and NHS staff survey 24b (knowing the system for reporting errors) (Table 16). In a reversal of results from the COPD criterion analysis there was a significant positive correlation between incidents reported to the NPSA and heart failure mean score, as the number of incidents increased the mean heart failure score increased (Figure 18). There was also a strong positive correlation between the percentage of responders to the NHS staff survey stating they did not know how to reporting errors confidentially and mean heart failure score. That is, as mean heart failure score increased the percentage of responders reporting that they did not know how to report concerns also increased (Figure 19).

**Table 16: Correlations between heart failure mean criterion score and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Incidents reported to NPSA per 100 bed days | 0.528 | 0.017 | Moderate |
| NHS staff survey Q24b: System for reporting (% no) | 0.620 | 0.004 | Strong |

|  |  |
| --- | --- |
| **Figure 18.**  **Scatter plot (with regression line) for heart failure criterion score by NPSA incidents bed days** | **Figure 19.**  **Scatter plot (with regression line) for heart failure criterion score by NHS staff survey Q24b: system for reporting (Percent no)** |
|  |  |

**Outliers in the analysis**

It was agreed to identify which hospitals were outliers in each analysis to establish whether it was the same trust each time. The following outliers were identified (hospital scores presented in brackets):

|  |  |
| --- | --- |
| COPD mean holistic overall score | Hospital RC9 (2.9) |
| HF mean holistic overall score | Hospital RN1 (3.1) |
| COPD mean holistic phase score | Hospital RQW (10.2) |
| HF mean holistic phase score | *No outlier identified* |
|  |  |
| COPD mean criterion based score | *No outlier identified* |
| HF mean criterion based score | Hospital RNH (9.3); Hospital RQW (7.0) |

Hospital RNH (from the heart failure criterion score analysis) was the most noticeable outlier in all analyses and a decision was taken to rerun the analysis to see if there were any changes in the correlation results.

There were some minor differences in results following the exclusion of hospital RNH. Incidents reported to the NPSA remained moderately correlated with heart failure mean score, where the higher the number of incidents the higher the mean heart failure score. The percentage of heart failure patients dying in hospital within 28 days moderately correlated with heart failure mean score: the lower the percentage of patients who die within 28 days the higher the mean score. Finally, NHS staff survey Q24b system for reporting error (% replying that they did not know of the system) is no longer strongly correlated with mean heart failure score.

***Regression Analysis***

A priori, one of the objectives of this analysis was to fit multiple regression models to COPD and heart failure scores to establish any predictors. With a sample of 20 hospitals the number of variables in a multiple regression model should be restricted to two as anything over two would result in over fitting a model. Furthermore, consideration should be taken of categorical variables such as existing national targets, where each level of response (not met, partly met, almost met, fully met) counts as one variable in the regression analysis – thus a categorical variable with three or more categories, which is the case for all categorical variables in this dataset, would result in over fitting.

It was possible to fit a multiple regression model for COPD criterion score where NPSA incidents and NHS staff survey question 22f (happy with standard of care provided) were significant predictors of COPD score at the 10% significance level. A one unit increase in the number of incidents reported to NPSA per 100 bed days resulted in a 0.5 decrease in COPD criterion score and a one unit increase in the staff happy with standard of care resulted in a 3.14 change in COPD score. Multiple regression models were fitted using the forwards stepwise method where the most significant variable at the univariate stage is selected first.

When the regression model was fitted to the COPD criterion score, this model only included the two variables that were significantly correlated with the score. When included in a regression model together ‘neither incidents per 100 bed days’ nor ‘NHS staff survey question 22f’ were significant at the 5% significance level (Table 17)

**Table 17: COPD criterion score regression with outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression coefficient  (standard error) | t-test | p-value |
| Constant | 0.82 (5.48) | 0.15 | 0.883 |
| Incidents reported to NPSA per 100 bed days | -0.50 (0.26) | -1.89 | 0.076 |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | 3.14 (1.75) | 1.79 | 0.091 |

A multiple regression model was also fitted to heart failure criterion score, where NPSA incidents and NHS staff survey question 22f were significant predictors of heart failure score at the 10% significance level. A one unit increase in the number of incidents reported to NPSA per 100 bed days resulted in a 0.21 increase in heart failure criterion score and a one unit increase in the percentage of staff reporting no to system for reporting errors resulted in a 0.11 change in heart failure score.

When a regression model was fitted to the heart failure criterion score, this model only included the two variables that were significantly correlated with criterion score. When included in a regression model together incidents per 100 bed days was no longer significant when included with NHS staff survey 24b (% of staff percent reporting they did not know of the system) (Table 18).

**Table 18: Heart failure criterion score regression with outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regression coefficient  (standard error) | t-test | p-value |
| Constant | 7.30 (0.19) | 38.98 | < 0.001 |
| Incidents reported to NPSA per 100 bed days | 0.21 (0.12) | 1.77 | 0.095 |
| NHS staff survey Q24b: System for reporting – percent reporting they did not know | 0.11 (0.04) | 2.56 | 0.020 |

For the remaining four analyses it was not possible to fit a multiple regression model, either because no two variables were significant in the same model or due to problems with over fitting for categorical variables.

**Conclusions**

This part of the study set out to explore how much of the variance in outcome for two important clinical conditions (COPD and heart failure) can be attributed to differences in quality of clinical care in acute hospitals. The main findings are:

* Within hospital and between hospital variation in quality of care can be identified from a combination of holistic scale scores and criterion based review.
* While there were trends towards hospitals having lower mortality also having higher quality of care scores, none of these differences were statistically significant
* Although there were some correlations between quality of care scores and hospital level outcome data, there was no clear relationship between the quality of care and hospital level outcomes for the two indicator conditions in this study. This may reflect the complexity of the process/outcome relationship at the patient group level.
* Available hospital level outcome indicator data are probably insufficiently sensitive to reflect the quality of care recorded in patient case notes. Furthermore, the nuances of patient care may mean that high quality care may be given even when the patient’s outcome appears poor, and vice versa. These findings may be pointing to process measures as being more useful than outcome measures when reviewing the case notes of people who have chronic disease or multiple conditions.

The study explored a complex methodological and clinical question and the extent to which it can be investigated is dependent on a number of factors, particularly in this study related to the available outcome measures and the availability of process of care measures. Choice of outcome measure can be critical in exploring the relationship between the process and outcome of care. For example, the death of a patient (outcome) in the terminal stages of a chronic illness may not be influenced even by the highest quality of care (process).

Previous studies have already indicated that the relationship between process and outcome is difficult to assess and the systematic review by Pitches et al[38] found that only about half of the 51 correlations between process measures and outcome indicators in the 36 studies were positive.

Process (or quality of care) measures in this study were derived from a multi-method case note review process that was refined during phase one of the study. For each case reviewers produced a synthesis of their perspective on the quality of care provided, rather than from direct observation. Only the quantitative data from holistic scale scores and criterion based scores could be used for the analysis, with the holistic textual data being used to validate the reviewer’s holistic scale scores being given to overall and phases of care.

Condition specific mortality and proxy outcomes were derived from a range of hospital level data. These measures were selected from a much more extensive list through a process of group discussion, drawing also on the experience in another study by the research team that examined the influence of hospital level process and outcome measures on incident reporting in acute hospitals.[53] Resource constraints and the difficulties of obtaining research governance and patient consent meant that, though desirable, it was not possible to obtain patient level outcome data for the 1565 cases reviewed.

In seeking to explore the relationship between the quality of care for the two tracer conditions in hospitals and overall quality of care and quality markers across hospital institutions, we found only a limited number of positive associations. Correlations with continuous outcome data, where positive (Table 1), mainly reflected the selection of mortality to group the hospitals.

In exploring the relationship between quality of care scores and hospital mortality group, the findings are consistent for both conditions. Across all phases of care and overall care using holistic review, and total criterion score, there was a trend towards higher mean scores for the lower mortality hospital groups (Tables 3 & 4). However, when account was taken of clustering by reviewer, there were no statistically significant differences in scores between the two groups of hospitals. These results are supported by the finding that there are no statistical differences for the mean holistic scale scores between the higher and lower mortality hospitals (Tables 5 – 8).

If the quality of care is good for one condition in a hospital as measured by case note review, should we expect that it will also be good for another condition? From each hospital there were approximately 40 reviews available for the quality of care in COPD and in heart failure (80 reviews in total). It might be hypothesised that a hospital from the lower mortality group might be expected to have higher quality of care scores– and that the levels of quality of care for the two conditions within each hospital bear some relationship. We did not find this – in general scores were only slightly higher in the lower mortality group than in the higher mortality hospital group. Analysis (Figure 3) shows that there is little association within hospitals and we are unable to show that better quality care results in better quality hospital level outcomes.

We explored the relationships between holistic and criterion based quality of care scores on the one hand and hospital level process and outcome indicators on the other. These indicators were drawn from a number of sources and their relationship with the outcome of quality of care for the two conditions took different forms. Some, such as the Healthcare Commission indicators, were related to the general external reference measures. Others, particularly from the NHS staff survey, are known to be related to safety culture and incident reporting.[53]

There are a number of positive correlations in the expected direction for holistic scale score and mean phase score data, although this is mainly for heart failure – there are only limited correlations for COPD holistic data. However there were also a number of negative correlations for which it is difficult to find an explanation. For example, as COPD mean holistic score increases the mean score for care of patient/service users being a top priority decreased. It is possible that some of these associations are chance findings because of the number of correlations being undertaken. However, we limited this effect by ensuring that all statistical tests were two-sided with a significance level of p ≤ 0.05 and weak and very weak associations were excluded from consideration in the analysis.

Why is the correlation between recorded process and outcome so apparently poor? We have already indicated that the methods may not be sophisticated enough to measure the associations. It may be that the use of hospital level measures that are not condition specific, and that reflect the organisation as a whole, are too abstract for the purpose of assessing care for people with such as COPD and heart failure.

However there may be other important confounding issues that relate to the meaning of quality of care and to the relationship between process and outcome when quality of care is measured.

When Mohamed and colleagues[27] explored the care of people with stroke they found, as we did, that both criterion and holistic methods are valuable in reviewing care and others have suggested similar approaches. They also found that clinical practice issues, such as the impact of Advance Directives, made it difficult to assess quality unless detailed information at the patient level was used to understand outcomes. In our study it became clear that reviewers could be very critical of quality of care and were willing to make explicit judgements about clinical practice. Crucially, though, reviewers were able to say that what might be regarded as a poor outcome (for example, a patient died) can be accompanied by very high quality of care. For a very ill patient, with little of no chance of surviving, high quality palliative care may be both appropriate and in the patient’s and family’s best interests. Alternatively, patients may survive very poor care. Again, choice of outcome measure is critical.

Nevertheless, although individual level data may increase the likelihood of defining process outcome relationships, this still remains a methodological challenge. When Gibbs and colleagues[7] used patient level as well as hospital level outcomes to investigate the relationship between process and outcome among surgical patients, they found that people who were more severely ill, or who died or had complications, had higher quality of care ratings than those with a lower predicted outcome risk. Furthermore, Pitches et al[38] have shown that numbers of studies find uncertain relationships between process and risk-adjusted outcomes of care. The results of this study seem to be following a similar trend.

*Limitations*

Because of a lack of proven methodologies for chronic disease care, and because of the difficulty of accessing individual outcome data, it was not possible to undertake a full risk adjusted analysis. Whereas Daley et al[37] were able to access a number of possible predictors for surgical care, these are mainly unavailable for chronic disease management. In this study the availability of the outcome measures was limited to hospital level data and it has not been possible to capture more individual level data. This is clearly a potential limitation. However, even though risk adjustment measures have been available for the assessment of the effect of process of care on surgical outcomes, a number of studies have found it difficult to show positive associations between risk adjusted mortality and good quality of care.[7][37][55][56]

Reviewers are not perfect. For instance, some may have a more positive view on cases than they should have, or be too inexperienced to identify flaws in care. Nevertheless we have no evidence that they glossed over errors since they did identify about 20% of cases which fell into the unsatisfactory range and about 4% that were identified as adverse incidents or near misses. The ability to judge the quality of care provided is also dependent on the quality of recording in the case notes and there may not be enough data in the records to be able to make an adequate judgement on quality of care – although reviewers indicated that about 85% of the case notes were in a satisfactory or better condition.

Although the two groups of hospitals in this study had considerable differences in mortality rates between the lower mortality rate hospitals and the higher mortality rate hospitals, a whole range of factors might account for those differences, such as case mix and age. Indeed, in this study, there is some indication that quality of care scores went down as age went up in the group of heart failure cases, but this may be because of higher levels of mortality risk before admission. We do not know enough about these cases to model this risk, though we did randomly select the 10 hospitals in each group.

We did not account for any measurement error in the process variables and the number of measurements per hospital. Since the correlation of true scores is equal to the correlation of observed scores divided by the square root of the product of their reliabilities it is possible that measurement error may dilute the magnitude of the correlation coefficients. This means that the correlations observed between the process and outcome variables may be lower than the true population correlation.

*Reviewing using a mixed method approach - how can it be used in future?*

Because of the research aims of the phase one study we separated out the review methods into two distinct sections - holistic and criterion based. In phase two those methods were combined to provide holistic scale scores, criterion based scores and textual data about the quality of care of each of the phases and of overall care. In this study the textual data was only used to validate the holistic scale score data. However, when reviewing or auditing small groups of cases, for example when there are concerns about the outcome of interventions, then mixed holistic and criterion based review, that also captures textual data, may prove a powerful model. In using these mixed methods careful attention will be required to aspects such as the selection and training of reviewers, including recognition of the problems associated with inter-rater reliability and bias.

**Overview of findings**

This study has not found strong relationships between quality of care, as measured by case note review, and outcome of care at the hospital level. Other authors have found a similar lack of direct association. The finding that reviewers considered quality of care as a broad concept, where patients who fared poorly overall because of their underlying condition might nevertheless have high quality care, reflects a similar finding by Gibbs et al[7] when examining patient level outcomes. In their study, patients with poor outcomes or with a higher predicted risk of mortality or morbidity had higher quality of care ratings than those with a low predicted risk of adverse outcome. Mohammed and colleagues[27] found that a combination of holistic data and criterion based data was required to understand the influence of care on outcomes and Pitches et al[38] concluded from their systematic review of the literature that ‘the general notion that hospitals with higher risk-adjusted mortality have poorer quality of care is neither consistent nor reliable’.

Lilford and colleagues[54] have recently provided a useful critique of risk adjusted outcomes in the assessment of healthcare quality in which they question the value of using outcome measures to evaluate the quality of care. We are unable to confirm that proposition with these data but would suggest that for the purposes of reviewing quality of care, process measures that allow for an integration both criterion based review and holistic review can provide a sound basis for decision making.

Meanwhile the real challenge of outcomes is to define what is measurable - and there are actually very few measures that are validated.[57] Where attempts have been made to develop outcome measures there has been some tendency for health professionals to dismiss outcomes as too difficult to use. But patients are concerned with the ultimate outcome of their therapy and thus we cannot ignore the need for appropriate measures, and risk adjustment methods, for chronic disease management. In the meantime, however, given the lack of agreement on specific outcome measures, we suggest that process measures are a reasonable proxy.

**Future research agenda**

1. Clinical staff are increasingly called upon to assess quality and safety of care from case notes under conditions where there is cause for concern. This research has identified aspects of case note review that could be used to support quality review and that would benefit from further research.

1.1. Research to assess the inter-rater reliability among experienced physician reviewers, including the effect of selection and training. This research should take account of other work from North America that has shown the potential for major discrepancies in inter-rater reliability.

1.2. Reviewer recruitment appears to be difficult in the UK, perhaps because this form of clinical review has been treated in an *ad hoc* fashion to date. A qualitative study should be undertaken to understand the possible barriers and factors that might enhance recruitment and training of clinician reviewers.

1.3. Doctors and nurses may view quality differently but there has been no UK research (and little international research) to explore whether experienced physicians and experienced specialist nurses review in similar or complimentary manners (possibly enhancing the overall scope and quality of reviews)

1.4. Research on the reliability of structured holistic assessment of the quality and safety of care using scales remains unusual in the literature. Evaluation of the sensitivity and internal consistency of these methods would be of value.

1.5. The extent to which review criteria are reproducible remains a research question. There is some evidence from US and UK studies using the RAND appropriateness method that panels constructing review criteria from the same clinical guideline have only moderate levels of agreement. The extensive need for review criteria generated by national clinical guidelines suggests further research would be useful before these criteria become a significant part of the new quality improvement programmes.

2. There is an important research agenda relating to linkage between process and outcome data for chronic disease care, and in relation to case note review. Although the potential for data retrieval from electronic records is considerable (information might be gathered from data mining and natural language programming), this agenda relates to paper based records which will remain the main data course in hospital care over the next five years in the UK.

2.1. There is a need for further research to explore risk adjusted outcome measure methodologies for chronic diseases. Methods of risk prediction should be developed in a fashion that enables the production of (at least some) measures across the spectrum of chronic disease, allowing better methods of outcome comparison.

2.2 There is a continued need for validated condition specific outcome measures for a range of chronic diseases, of a type that can be used in health services evaluation – that is, comprising a minimum data set of items that might, in future evaluations, be collectable through electronic records systems.

We thank the external referees for their contributions to shaping this research agenda.

**References**

1] Codman EA. *A Study in Hospital Efficiency. As Demonstrated by the Case Report of the First Five Years of a Private Hospital*. Boston: Thomas Todd Co., 1918-1920.

2] Rubenstein LV, Kahn KL, Reinisch EJ, Sherwood MJ, Rogers WH, Kambers C. Changes in quality of care for five diseases measured by implicit review, 1981-1986. *JAMA*, 1990;**264**:1974-1979.

3] Leape LL, Brennan TA, Laird N, Cawthers AG, Localio AR, Barnes BA. et al. The nature of adverse events in hospitalised patients: results of Harvard Medical Practice Study.II. *NEJM*, 1991;**324**:377- 384.

4] Runciman WB, Webb RK, Helps SC, Thomas EJ, Sexton EJ, Studdert DM, Brennan TA. A comparison of iatrogenic injury studies in Australia and the USA II reviewer behaviour and quality of care. *Int J Quality in Health Care*, 2000;**12**(5):379-388.

5] Thomas EJ, Studdert DM, Burstein HR, Orav JE, Zeena TBS, Williams EJ, Howard KM, Weiler PC, Brennan TA. Incidences and types of adverse events and negligent care in Utah and Colorado. *Medical Care*, 2000;**38**(3).261-271.

6] Weingart SN, Davis RB, Palmer RH, Cahalane M, Hamel MB, Mukamal K, Phillips RS, Davies DT, Iezzoni LI. Discrepancies between explicit and implicit review: Physician and Nurse Assessments of Complication and Quality. *Health Services Research*, 2002;**32**:483- 498.

7] Gibbs J, Clark K, Khuri S, Henderson W, Hur K, Daley J. Validating risk adjusted surgical outcomes: chart review of process of care. *Int J Qual Health Care,* 2001:**13**(3);187-196.

8] Ashton C, Kuykendall D, Johnson ML, Wray N. An empirical assessment of the validity of explicit and implicit process of care criteria for quality assessment, *Med Care*,1999;**37**:798-808.

9] Thomas EJ, Studdert DM, Brennan TA. The reliability of medical record review for estimating adverse event rates. *Ann Int Med*, 2002;**136**(11)812-816.

10] Hofer TP, Asch SM, Hayward RA, Rubenstein LV, Hogan MM, Adams J, Kerr EA. Profiling quality of care: is there a role for peer review? *BMC Health Services Research* 2004;**4**:9. <http://www.biomedcentral.com/1472-6963/4/9>

11] Hayward RA, Hofer TPE. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. *JAMA*, 2001;**286**:415-420.

12] Lilford R, Edwards A, Girling A, Hofer T, Di Tanna GL, Petty J, Nicholl J. Inter-rater reliability of case-note audit: a systematic review. *J Health Serv Res Policy*, 2007;**12**:173-180

13] Hulka BS, Romm FJ, Parkerson GR, Russell IT, Clapp NE, Johnson FS. Peer review in Ambulatory Care: Use of Explicit Criteria and Implicit Judgements. *Med Care*, 1979;**17**(Suppl):1-73.

14] Hayward RA, McMahon LF, Bernard AM. Evaluating the care of general medicine inpatients: how good is implicit review? *Ann Intern Med*, 1993;**118**:550-556

15] Fischoff B. Hindsight ≠ foresight: The effect of outcome knowledge on judgement under uncertainty. *J Experimental Psychol*, 1975;**1**:288-299

16] Lilford RJ, Mohammed MA, Brauholtz D, Hofer TP. The measurement of active errors: methodology issues. *QSHC*, 2003;**12**(SuppI):ii8-ii12

17] Localio RA, Weaver SL, Landis R, Lawthers AG, Brennan TA, Hebert L, Sharp TJ. Identifying adverse events caused by medical care: degree of physician agreement in a retrospective chart review. *Ann Intern Med*, 1996;**125**:457-464.

18] Agency for Health Care Policy and Research. *Using clinical practice guidelines to evaluate quality of care. Volume 2.* Rockville MD, US Department of Health and Human Services, Public Health Service, 1995.

19] Hadorn DC, Baker DW, Kamberg CJ, Brook RH. Practice guidelines. Phase II of the AHCPR – sponsored heart failure guideline: translating practice recommendations into review criteria. *J Quality Improvement*, 1996;**22**:265-276.

20] The North of England Study of Standards and Performance in General Practice. Medical audit in general practice. I: Effects on doctors' clinical behaviour for common childhood conditions. *BMJ,* 1992;**304:**1480-1484.

21] Hutchinson A, McIntosh A, Anderson J, Gilbert C, Field R. Developing primary care review criteria from evidenced-based guidelines: coronary heart disease as a model. *BJGP*, 2003;**53**:691-696.

22] National Institute for Clinical Excellence. *Chronic Heart Failure: management of chronic heart failure in adults in primary and secondary care. Clinical Guideline 5.* London, National Institute for Clinical Excellence, 2003

23] Rudd AG, Lowe D, Irwin P, Rutledge Z, Pearson M; Intercollegiate Stroke Working Party. National stroke audit: a tool for change? *Qual Health Care*. 2001 Sep;**10**:141-51.

24] Gompertz P, Dennis M, Hopkins A, Ebrahim S. Development and reliability of the stroke audit form. UK Stroke Audit Group. *Age Aging*, 1994;**23**:378-83.

25] Gompertz PH, Irwin P, Morris R, Lowe D, Rutledge Z, Rudd AG, Pearson MG. Reliability and validity of the Intercollegiate Stroke Audit Package. *J* *Eval Clin Prac*, 2001;**7**:1-11.

26] Camacho LA, Rubin HR. Assessment of the validity and reliability of three systems of medical record screening for quality of care assessment. *Med Care*, 1998;**36**:748-51.

27] Mohammed MA, Mant J, Bentham L, Stevens A, Hussain S. Process and mortality of stroke patients with and without do not resuscitate order in the West Midlands, UK. *Int J Qual Healthcare*, 2006;**18**:102-106.

28] Rubenstein LR, Kahn KL, Harris ER, Sherwood MJ, Rodgers WH, Brook RH. *Structured implicit review of the medical record: a method for measuring the quality of inhospital medical care and a summary of quality changes following implementation of the Medicare prospective payments system*. Santa Monica, RAND, 1991

29] Pearson M, Lee JL, Chang BL, Elliott M, Kahn KL, Rubenstein LV. Structured implicit review: a new method for monitoring nursing care quality. *Med Care*, 2000;**38**:1074-1091.

30] Keeler EB, Rubenstein KLK, Draper D, Harrison ER, McGinty MJ, Rogers WH, Brook RH, Health Programme of RAND. *JAMA*, 1992;**268**:1702-1708.

31] National Institute for Clinical Excellence. *Chronic obstructive pulmonary disease. Management of chronic obstructive pulmonary disease in adults in primary and secondary care.* London, National Institute for Clinical Excellence, 2004

32] Van Belle, Gerald. *Statistical Rules of Thumb. Wiley Series in Probability and Statistics.* Wiley-Interscience; 1st edition, 2002

33] Hospital Episode Statistics Online: Data on hospital providers.

[www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=212](http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=212) (accessed 20 November 2006).

34] Royal College of Physicians and British Thoracic Society. Report of the 2003 National COPD Audit. London: Royal College of Physicians, 2004.

35] Fleiss JL. *Statistical methods for rates and proportions, 2nd Edition*. New York, Wiley, 1981

36] Dale, JR. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 1986;**42:**909-917.

37] Daley J, Khuri SF, Henderson W, Hur K, Gibbs J, Barbour G, Demakis J, Irvin G, Stremple J, Grover F, MCdonald G, Passaro E, Fabri P, Spencer J, Hammermeister K, Aust J, Oprian C. Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: results of the National Veterans Affairs surgical risk study. *J Am Coll Surg.* 1997;**185**:315-327.

38] Pitches DW, Mohammed MA, Lilford RJ. What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Services Research* 2007;**7**:91 <http://www.biomedcentral.com/1472-6963/7/91>

39] Glaser, B. and Strauss, A.. *Discovery of grounded theory: strategies for qualitative research.* 1967. London: Weidenfeld and Nicolson.

40] [http://www.nhsemployers.org/restricted/downloads/download.asp?ref=363&hash= 0bb7dc2313394a93337d3adf51cf6c3f&itemplate=e\_aboutus\_3col\_aboutus-2028](http://www.nhsemployers.org/restricted/downloads/download.asp?ref=363&hash=%200bb7dc2313394a93337d3adf51cf6c3f&itemplate=e_aboutus_3col_aboutus-2028)

41] Wardle TD, Burnham R, Greig E, Preston S, Harris RA, Borrill Z, Driscoll P, Seward E, Harrison BDW, Lowe D, Pearson MG.   A confidential study of deaths after emergency medical admission: Issues relating to quality of care.   *Clin Med* 2003;**3**:425-34.

42] Potter J, Peel P, Mian S, Lowe D, Irwin P, Pearson M, Wagg A. National audit of continence care for older people: management of faecal incontinence. *Age Ageing*, 2007;**36**:268-73.

43] Hutchinson A, McIntosh A, Coster JE, Cooper KL, Bath PA, Walters SJ, Pearson M. *When is an event an event? Data from the quality and safety continuum. In: From safe design to safe practice.* Eds. Hignett S, Norris B, Catchpole K, Hutchinson A, Tapley S. Cambridge, The Ergonomics Society, 2008. ISBN 978-0-9554225-2-2

44] Armitage, P, Berry, G, and Matthews, J.N.S. *Statistical Methods in Medical Research.* 4th edition. Blackwell Science, Oxford, 2002.

45] Luck J, Peabody JW, Dressellhaus TR, Lee M, Glassman P. How well does chart abstraction measure quality? A prospective comparison of standardised patients with the medical record. *Am J Med* 2000;**108**:642-649

46] Healthcare Commission.

[www.healthcarecommission.org.uk](http://www.healthcarecommission.org.uk/nationalfindings/surveys.cfm) (accessed 30 November 2008).

47] Intercollegiate Stroke Audit Working Party. *National Sentinel Stroke Audit*. London: Royal College of Physicians of London, 2007

48] McNaughton H, McPherson K, Taylor W, Weatherall M. Relationship between process and outcome in stroke care. *Stroke*, 2003:**34**;713-717.

49] Wilson B, Thornton JG, Hewison J, Lilford RJ, Watt I, Braunholtz D, Robinson M. The Leeds University maternity audit project. *Int J Qual in Health Care*, 2002;**14**:175-181.

50] Dr Foster. The Hospital Guide: How Good is My Hospital, 2005. [www.drfoster.co.uk/hospitalreport/pdfs/howGood.pdf](http://www.drfoster.co.uk/hospitalreport/pdfs/howGood.pdf) (accessed 20 November 2006).

51] Healthcare Commission NHS Staff Survey.

[www.healthcarecommission.org.uk/nationalfindings/surveys.cfm](http://www.healthcarecommission.org.uk/nationalfindings/surveys.cfm) (accessed 20 November 2006).

52] National Patient Safety Agency. Quarterly National Reporting and Learning System Data Summary, Autumn 2006.

[www.npsa.nhs.uk/site/media/documents/1953\_NRLS\_Data.pdf](http://www.npsa.nhs.uk/site/media/documents/1953_NRLS_Data.pdf) (accessed 20 November 2006).

53] Hutchinson A, Young TA, Cooper KL, McIntosh A, Karnon JD, Scobie S, Thomson RG. Trends in healthcare incident reporting and relationship to safety and quality data in acute hospitals: results from the National Reporting and Learning System. *Qual Saf Health Care* 2009;**18**:5-10

54] Lilford RJ, Brown CA, Nicholl, J. Use of process measures to monitor the quality of clinical practice. *BMJ*, 2007;**335**:648-650.

55] Dubois RW, Rogers WH, Moxley JH et al. Hospital inpatient mortality: is it a predictor of quality? *N Engl J Med*, 1987;**317**:1674-1680

56] Thomas JW, Holloway JJ, Guire KE. Validating risk-adjusted mortality as an indicator of quality of care. *Inquiry* 1993;**30**:6-22

57] Pearson M, Goldacre M, Coles J, Amess M, Cleary R, Fletcher J, Mason A, Dixon P, Eastwood A. *Health Outcome Indicators: Asthma Report of working group to the Department of Health*. Oxford: National Centre for Health Outcomes Development, 1999.

**Appendix 1**

**COPD review criteria**

**NICE definition of an exacerbation of COPD:**

ICD-10: J43, 44

*“An exacerbation is a sustained worsening of the patient’s symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, cough, increased sputum production and change in sputum colour.” [32]*

**Key data**

1. Audit record number

2. Hospital number

3. Date of birth

4. Gender

male

female

5. First part of patient’s postcode

6. Date of this admission to hospital

7. Did the patient die during this admission

yes

no

If yes, was the recorded cause of death

COPD or complications of COPD

other cause(s)

not recorded

8. Date of discharge from hospital (or death if applicable)

9. Was the patient accepted by an early discharge  
(or hospital at home) scheme? yes

no

not applicable

10. Prior to this admission, has the patient previously been admitted  
to hospital for COPD or accepted onto an early discharge scheme?

yes

no

**History/patient characteristics**

11. What is the patient’s smoking status?

current smoker

ex smoker (stopped more than 3 months)

life long non-smoker

not recorded

If current or ex-smoker,

how many cigarettes smoked per day?

***or***

pack years?

don’t know

12. Does the patient have any significant co-morbidities?

Please tick all that apply

none

heart disease   
 hypertension

stroke

locomotor problems

neurological problems

diabetes

visual impairment

depression/anxiety

other

13. What are the patient’s social circumstances?

lives alone, no support

lives alone with social service support

lives with spouse or close relative

lives in nursing/residential home

lives in warden controlled (sheltered) housing

**Admission**

14. At admission, was there a record of

level of breathlessness increased

not increased

not recorded

level of sputum increased

not increased

not recorded

changes in colour of sputum changed

not changed

not recorded

sputum colour white or grey

yellow or green

no sputum

not recorded

15. Was the patient’s dyspnoea rating (e.g. on the MRC dyspnoea scale) recorded

yes

no

16. What is the patient’s performance status?

normal activity

strenuous activity limited

limited activity but self care

limited self care

bed/chair bound, no self care

not known

17. Was a chest x-ray taken? yes

no

If yes, is the x-ray report in the notes? yes

no

18. Was respiratory rate measured?

yes

no

If yes, what was the first reading  
after admission per minute

19. Were blood gases taken? yes

no

If yes, what was the first recorded (after admission) value for?

pH or: H+ (mmol/l) not recorded

PCO2 (kPa) or: mmHg not recorded

PO2 (kPa) or: mmHg not recorded

Was level of O2 to be given stipulated in notes/on chart? yes

no

20. Was an ECG performed? yes

no

21. Was urea recorded? yes

no

If yes, what was the first recorded (after admission) value?

mmol/l

not recorded

22. Was serum albumin recorded? yes

no

If yes, what was the first recorded (after admission) value?

mmol/l

not recorded

23. Was there a record of medications being taken at time of admission?

yes

no

If yes, were there 5 or more medications recorded?

yes

no

24. What was the patient’s temperature at admission?

not recorded

25. Is there a spirometry reading in the notes for this admission?

Yes

No

If yes what is the FEV1 level

(if more than one, give most recent)

not recorded

26. Is there a record of peripheral oedema?

yes - present

yes – not present

not recorded

If present, was it?

leg/ ankles

sacral

not recorded

**Initial management**

27. Was a course of antibiotics prescribed?

yes

no

28. Were nebulised bronchodilators prescribed?

yes

no

29. Did the patient receive systemic corticosteroids?

yes

no

30. How many sets of arterial blood gases results are in the records for this stay?

31. Did the patient have a pH less than 7.35 at any time during this stay?

yes

no

If yes, did they receive ventilatory support?

respiratory stimulant (e.g. doxapram)

non-invasive

invasive

none

If the patient had a pH of less than 7.35 and did not receive ventilatory support,  
is it noted why not?

patient refused

no facilities

not appropriate

failed

other (please state)

not recorded

**Pre-discharge phase**

32. Was oximetry (O2 saturation levels) undertaken, after acute phase but prior to discharge?

yes

no

not recorded

If yes, what were the results? % not recorded

33. If a current smoker, was help toward smoking cessation given

referred to smoking cessation programme

advice given and recorded

nothing recorded

not applicable (because non-smoker)

34. Was there an assessment of the patient’s home circumstances and their ability to cope?

yes

no

not recorded

35. Where was the patient discharged to?

own home – independent of help

own home – with additional social support

sheltered housing or living with relative

nursing or residential care

not applicable – died in hospital

36. Is there a letter to the patient’s primary care team?

yes

no

not recorded

If yes, did the letter include a clear list of the patient’s medication?

yes

no

not recorded

37. Which type of consultant was the patient under at  
time of discharge?

respiratory physician

care of elderly physician

general physician

other

not recorded

**Appendix 2**

**Heart failure review criteria**

*NICE definition of an exacerbation of Heart Failure: (Heart failure due to left ventricular systolic dysfunction)*

*ICD –10 - I50.0, I50.1, I50.9, I11.0*

"An exacerbation of heart failure is a sustained worsening of the patient's symptoms from their usual stable state which is beyond normal day-to-day variations, and is acute in onset. Commonly reported symptoms are worsening breathlessness, tiredness and swelling of the feet and/or ankles." [23]

**Key data**

1. Audit record number

2. Hospital number

3. Date of birth

4. Gender

male

female

5. First part of patient’s postcode

6. Date of admission to hospital

7. Did the patient die during this admission

yes

no

If yes, was the recorded cause of death

heart failure or complications of heart failure

other cause(s)

not recorded

8. Date of discharge from hospital (or death if applicable)

**Patient history**

9. What is the patient’s smoking status current smoker

ex smoker (stopped more than 3 months)

life long non-smoker

not recorded

If current or ex-smoker,

how many cigarettes smoked per day?

or

pack years?

don’t know

10. Is there a record of the patient’s weekly alcohol intake?

units per week

not recorded

11. Does the patient have any significant co-morbdities?

Please tick all that apply

none

hypertension

stroke

respiratory disease

locomotor problems

neurological problems

visual impairment

diabetes

depression/anxiety

other

**Investigations/examination**

12. Was an electrocardiogram performed? yes

no

13. Was an echocardiogram performed during this stay? yes

no

Is it stated in the notes that an echocardiogram  
was previously performed

in the past year

ever

not recorded

Is it stated that an echocardiogram showed

normal left ventricular (LV) function

abnormal left ventricular (LV)function

not recorded

14. Have peptides (ANP or BNP) been measured? Yes

No

15. Was a spirometry or other pulmonary function reading taken? Yes

No

16. Was a chest x-ray performed? yes

no

not recorded

If yes, is the x-ray report in the notes? yes

no

not recorded

17. What was the first recorded (after admission) value for

urea mg/dl / mmol/l

not recorded

glucose mg/dl / mmol/l

not recorded

18. Is there a record for any of the following used to assess functional capacity?

Tick all that apply:

breathlessness scale (eg NYHA scale)

objective measure of O2 capacity

treadmill test without O2 consumption test

cardiopulmonary exercise test with Os consumption

none recorded

19. Is there a record of raised jugular venous pressure  
(also called jugular venous distension)

yes – elevated

yes – not elevated

not recorded

20. Is there a record of peripheral oedema? yes - present

yes – not present

not recorded

If present, was it?

leg/swollen ankles

sacral

not recorded

21. What was first blood pressure (on arrival)?

mmHg

not recorded

**Initial management**

22. For each of the following medications, is it recorded if the patient was prescribed

diuretic

prior to admission

prescribed during stay

on discharge

not prescribed

not recorded

ACE inhibitor

prior to admission

prescribed during stay

on discharge

not prescribed

not recorded

beta blocker

prior to admission

prescribed during stay

on discharge

not prescribed

not recorded

23. Is it recorded that the patient received oxygen during this stay?

yes

no

24. Is there a record of the patient’s weight in the notes?

yes

no

25. Is there a record of the patient’s blood pressure (daily) in the notes?

yes

no

26. Is there a record of the patient’s electrolytes in the notes?

yes

no

**Pre-discharge phase**

27. Pre-discharge assessments

Was blood pressure taken prior to discharge (within 48 hours of discharge)?

yes

no

not recorded

Was the level of breathlessness prior to discharge (within 48 hours of discharge) documented?

breathless at rest

breathless on minor exertion

breathless on strenuous exercise

not recorded

Was patient’s weight prior to discharge (within 48 hours of discharge) recorded?

yes

no

Was serum creatinine recorded within 48 hours of discharge

yes

no

28. Is there a past record of/ or plan to refer for

an exercise/rehabilitation programme

yes

no

29. If a current smoker, was help toward smoking cessation given

referred to smoking cessation programme

advice given and recorded

nothing recorded

not applicable (because non-smoker)

30. Was there an assessment of the patient’s home circumstances and their ability to cope?

yes

no

31. Where was the patient discharged to?

own home – independent of help

own home – with additional social support

sheltered housing or living with relative

nursing or residential care

not applicable – died in hospital

32. Is there a letter to the patient’s primary care team?

yes

no

If yes, did the letter include a clear list of the patient’s medication?

yes

no

not recorded

33. Which type of consultant was the patient under at  
time of discharge?

cardiologist

care of elderly physician

general physician

other

not recorded

**Appendix 3**

**Validity of review criterion questionnaire (COPD)**

**Key data**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No** | **Criterion** | **Essential** | **Desirable** | **Non-Essential** | **Comments** |
| 1 | Audit record number  (RReSQ Study Reference Number) |  |  |  |  |
| 2 | *Hospital number* |  |  |  |  |
| 3 | Date of birth |  |  |  |  |
| 4 | Gender |  |  |  |  |
| 5 | *First part of patient’s postcode* |  |  |  |  |
| 6 | Date of this admission to hospital (dd/mm/yyyy) |  |  |  |  |
| 7 | Did the patient die during this admission?  yes  no |  |  |  |  |
|  | If yes, was the recorded cause of death  COPD or complications of COPD  other cause(s)  not recorded |  |  |  |  |
| 8 | Date of discharge from hospital (or death if applicable)  discharge  death |  |  |  |  |
| 9 | Was the patient accepted by an early discharge (or hospital at home) scheme? yes  no  not applicable |  |  |  |  |
| 10 | Prior to this admission, has the patient previously been admitted to hospital for COPD or accepted onto an early discharge scheme?  yes  no |  |  |  |  |

**History/patient characteristics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Criterion** | **Essential** | **Desirable** | **Non-Essential** |
| 11 | What is the patient’s smoking status?  current smoker  ex smoker (stopped more than 3 months)  life long non-smoker  not recorded |  |  |  |
|  | If current or ex- cigarette smoker,  how many cigarettes smoked per day?  ***or***  pack years?  don’t know |  |  |  |
| 12 | Does the patient have any co-morbidities?  Please tick all that apply  none  heart disease  hypertension  stroke  locomotor problems  neurological problems  diabetes  visual impairment  depression/anxiety  other |  |  |  |
| 13 | What are the patient’s social circumstances?  lives alone, no support  lives alone with social service support  lives with spouse, close relative or carer  lives in nursing/residential home  lives in warden controlled (sheltered) housing  not known |  |  |  |

**Admission**

| **No** | **Criterion** | **Essential** | **Desirable** | **Non-Essential** |
| --- | --- | --- | --- | --- |
| 14 | At admission, was there a record of  level of breathlessness  increased  not increased  not recorded |  |  |  |
|  | level of sputum  increased  not increased  not recorded |  |  |  |
|  | changes in colour of sputum  changed  not changed  not recorded |  |  |  |
|  | sputum colour  white or grey  yellow or green  no sputum  not recorded |  |  |  |
| 15 | Was the patient’s dyspnoea rating (e.g. on the MRC dyspnoea scale) recorded  yes  no |  |  |  |
| 16 | What is the patient’s performance status, prior to admission?  normal activity  strenuous activity limited  limited activity but self care  limited self care  bed/chair bound, no self care  not known |  |  |  |
| 17 | Was a chest x-ray taken within 24 hours?  yes  no |  |  |  |
|  | If yes, is the x-ray report in the notes?  yes  no |  |  |  |
| 18 | Was respiratory rate measured within 24 hours?  yes  no |  |  |  |
|  | If yes, what was the first reading after admission (per minute) |  |  |  |
| 19 | Were blood gases taken within 24 hours?  yes  no |  |  |  |
|  | If yes, what was the first recorded (after admission) value for?  pH or: H+ (mmol/l)  not recorded  PCO2 (kPa) or: mmHg  not recorded  PO2 (kPa) or: mmHg  not recorded |  |  |  |
|  | Was level of O2 to be given stipulated in notes/on chart?  yes  no |  |  |  |
| 20 | Was an ECG performed?  yes  no |  |  |  |
| 21 | Was urea recorded?  yes  no |  |  |  |
|  | If yes, what was the first recorded (after admission) value?  mmol/l  not recorded |  |  |  |
| 22 | Was serum albumin recorded?  Yes  no |  |  |  |
|  | If yes, what was the first recorded (after admission) value?  mmol/l  not recorded |  |  |  |
| 23 | Was there a record of medications being taken at time of admission (within 24 hours)?  yes  no |  |  |  |
|  | If yes, were there 5 or more medications recorded?  Yes  no |  |  |  |
| 24 | What was the patient’s temperature at admission? (oC)  not recorded |  |  |  |
| 25 | Is there a spirometry reading in the notes for this admission?  Yes  no |  |  |  |
|  | If yes what is the FEV1 level (if more than one, give most recent)  not recorded |  |  |  |
| 26 | Is there a record of peripheral oedema?  yes - present  yes - not present  not recorded |  |  |  |
|  | If peripheral oedema was present, was it?  leg/ankles  sacral  not recorded |  |  |  |

**Initial management**

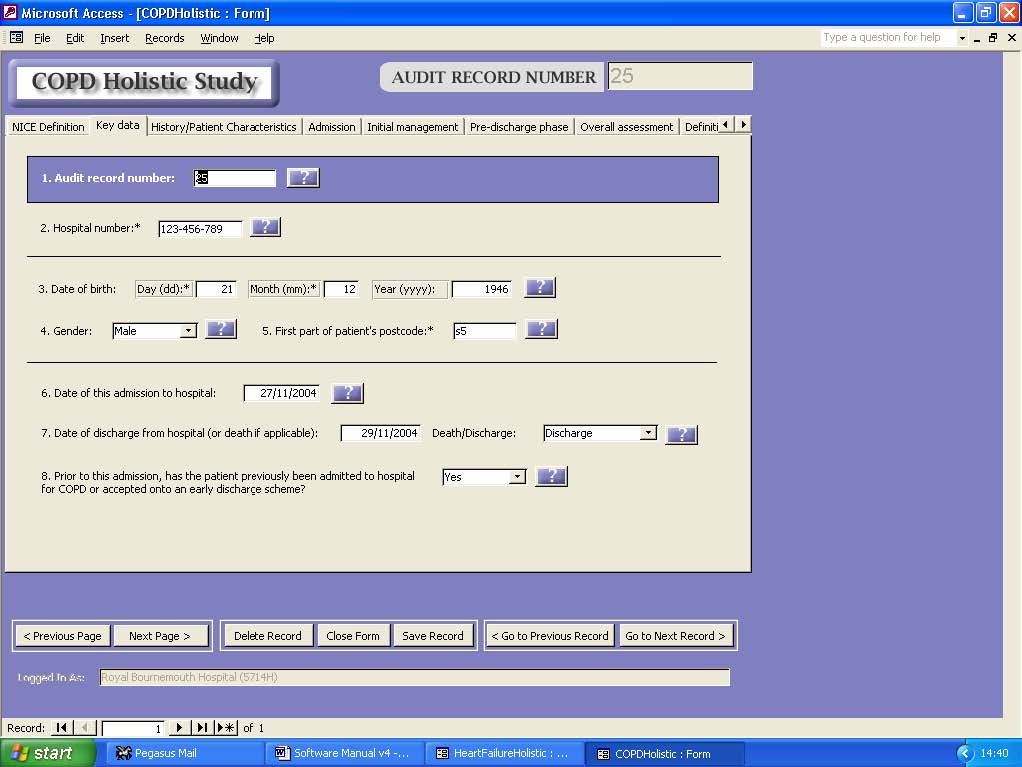
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Criterion** | **Essential** | **Desirable** | **Non-Essential** |
| 27 | Was a course of antibiotics prescribed?  yes  no |  |  |  |
| 28 | Were nebulised bronchodilators prescribed?  yes  no |  |  |  |
| 29 | Did the patient receive systemic corticosteroids?  yes  no |  |  |  |
| 30 | How many sets of arterial blood gases results are in the records for this stay? |  |  |  |
| 31 | Did the patient have a pH less than 7.35 at any time during this stay?  yes  no |  |  |  |
|  | If yes, did they receive ventilatory support?  Please tick all that apply  respiratory stimulant (e.g. doxapram)  non-invasive  invasive  none |  |  |  |
|  | If the patient had a pH of less than 7.35 and did not receive ventilatory support, is it noted why not?  patient refused  no facilities  not appropriate  failed  other  not recorded |  |  |  |

**Pre-discharge phase**

| **No** | **Criterion** | **Essential** | **Desirable** | **Non-Essential** |
| --- | --- | --- | --- | --- |
| 32 | Was oximetry (O2 saturation levels) undertaken, after acute phase but prior to discharge  (within 48 hours of discharge)?  yes  no |  |  |  |
|  | If yes, what were the results?  not recorded |  |  |  |
| 33 | If a current smoker, was help toward smoking cessation given  referred to smoking cessation programme  advice given and recorded  nothing recorded  not applicable (because non-smoker) |  |  |  |
| 34 | Was there an assessment of the patient’s home circumstances and their ability to cope?  yes  no |  |  |  |
| 35 | Where was the patient discharged to?  own home – independent of help  own home – with additional social support  sheltered housing or living with relative/carer  nursing or residential care  other hospital  not applicable – died in hospital  not recorded |  |  |  |
| 36 | Is there a letter to the patient’s primary care team?  yes  no |  |  |  |
|  | If yes, did the letter include a clear list of the patient’s medication?  yes  no |  |  |  |
| 37 | Which type of consultant was the patient under at time of discharge?  respiratory physician  care of elderly physician  general physician  other  not recorded |  |  |  |
| 38 | Time taken to complete (in hours and minutes) |  |  |  |

**Appendix 4**

**Screen shot of a holistic review data collection page**



**Microsoft Access ©**

**Appendix 5**

**Reviewer training scenarios to assist in recognising variation in care quality using holistic review**

The scenarios used stroke care as exemplars in order not to influence the views of the reviewers on what they might perceive as appropriate care for either of the two tracer conditions, COPD or heart failure.

**Mrs X scenario 1**: Mrs X, a 78 year old lively lady, has a dizzy turn and for an hour looses the use of her right arm and feels weakness in her right leg. She has had a TIA and goes to her GP very concerned. He gets her an outpatient appointment at her local hospital for two months’ time. A week after she has seen him she suffers a completed stroke (involving her right side and speech) and is taken to hospital where she is admitted to a general medical ward. It is a Friday. She is seen by a consultant on the following Tuesday. She has a CT scan on the Thursday. She develops pneumonia on the Friday, which is treated. She is referred the following week to a geriatrician. She is eventually sent home very disabled, no information is sent to her GP, and there proves to be a long wait for community rehabilitation, so after 2 weeks she is sent to a nursing home where she dies after 3 months.

Quality comments:

This lady showed early signs of a pending stroke, which could have been prevented.

* No early diagnosis is made and no rapid referral to a specialist service is provided
* No aspirin is prescribed to prevent the stroke.
* No specialist care in a stroke unit
* There is a delay in providing a CT scan
* Dysphagia (difficulty swallowing) is not picked up by early screening, leading to pneumonia
* She does not receive early therapy input to begin rehabilitation and mobilisation
* The communication systems between hospital and community fail.
  + The community services are unable to provide the rehabilitation she needs.

**Mrs X scenario 2**: Mrs X, a 78 year old lively lady, has a dizzy turn and for an hour looses the use of her right arm and feels weakness in her right leg. She has had a TIA and goes to her GP very concerned. He gets her an appointment for the TIA clinic in 2 days’ time and starts her on aspirin. At the TIA one-stop clinic she is duplex scanned and referred for early carotid endarterectomy, which she has within 10 days. She returns home on appropriate secondary prevention, having received dietary advice to eat more healthily. She was a life-long smoker so has been given cessation advice and nicotine replacement therapy. She continues with her active life, with regular follow-up visits to her GP.

Quality comments:

Evidence shows the danger of stroke is high after a TIA.

* + She receives appropriate rapid referral to a specialist service
  + Secondary prevention with aspirin is commenced immediately.
  + Early investigations are carried out
  + Appropriate surgical treatment is provided quickly
  + Secondary prevention and lifestyle advice are provided prior to discharge.
* Her health is monitored regularly by her GP.

**Mr Y scenario 3**: Mr Y is a 60 year old man. He has a rapid onset stroke on a Saturday and is taken to his local hospital. Because the hospital CT scanner does not work over the weekend, he is booked for a scan later in the week but this is cancelled for emergency admissions and somehow never gets re-booked. There is no acute stroke unit and, the rehab stroke unit being full, he is sent to the acute medical admissions ward. There he is not screened for dysphagia so his problems with swallowing are not picked up and no protocol begins. He also develops pneumonia after a week. Because of his incontinence he is catheterised in the AMU and the catheter remains in situ when he is sent to the general medical ward and somehow gets left there. He is visited by the physiotherapist for his pneumonia and begins mobilising half an hour a day when his pneumonia has resolved. Mobilising is also made more difficult by his catheter and the fact he feels so weak from not eating properly and following the pneumonia. Other complications arise from his catheter. He spends many weeks in hospital and eventually is sent home, catheter still in situ because he has persisting incontinence from poor bladder tone he developed from prolonged catheterisation. He remains weak and increasingly disabled. No rehabilitation continues at home and after a year he suffers a second stroke from which he does not recover. The family feel he had poor care and are making a complaint through the Healthcare Commission.

Quality comments:

* A CT scan is never carried out
* He does not receive specialist care in a stroke unit
* His dysphagia (difficulty swallowing) is not picked up, leading to pneumonia and malnutrition
* Prolonged, possibly unnecessary, use of a catheter leads to complications
* No early supported discharge from hospital is offered
* No rehabilitation is provided on his return to primary care.

**Mr Y scenario 4**: Mr Y is a 60 year old man. He has a rapid onset stroke on a Saturday and is taken to his local hospital. There, despite it being a weekend, he has a CT scan where ischaemic stroke is confirmed. He is admitted directly to the acute stroke unit. He is attached to monitoring systems (oxygen levels and ECG) by specialist nurses who take regular observations of his vital signs (blood sugars, temperature, etc). His clinical assessment on arrival in the unit, which includes a swallow screen, identifies he has dysphagia. He is made nil by mouth and tube feeding is begun by the nurses using the protocol agreed with the stroke team. The speech and language therapist sees him on the Monday to begin treatment and his dietary regime comes under the guidance of the specialist dietician. He is incontinent of urine in the first 48 hours which is conservatively managed without catheterisation and resolves spontaneously. The physiotherapist begins his early mobilisation regime which the nurses practice with him. An OT referral is made and therapy begins within one week of referral. After a fortnight the secondary prevention regime is established (aspirin having been started on the second day) using a protocol based on the national clinical guidelines for stroke. Because Mr Y is anxious to get home by this stage he has been referred to the specialist stroke Early Supported Discharge team and, once it is clear he can safely get himself out of bed, he goes home under their care for continuing rehabilitation at home.

Quality comments:

* A CT scan is carried out early, confirming the diagnosis
* Secondary prevention with aspirin is commenced early
* Hospital care is provided in a specialised stroke unit
* He receives regular monitoring of physiological indicators while in hospital
* Dysphagia is identified early and managed, reducing the risk of pneumonia and malnutrition
* Urinary incontinence is managed without the need for a catheter (this is associated with better clinical outcome)
* Mobilisation therapy is begun early and monitored (this is one of the key features of stroke units associated with better clinical outcome)
* Early supported discharge by a specialist team is deemed suitable and is provided (this is proven to result in the same outcomes as hospital rehabilitation and is liked by patients).

**Appendix 6**

# Record Review for Safety and Quality Study

*A report of the case note reviews undertaken*

*by reviewer xxxx*

**A collaborative project by the Royal College of Physicians and the University of Sheffield**

**FOREWORD**

Medical record review has become a standard means of assessing variations in quality of care. This is despite uncertainty about which methods of record review are most effective and reliable. The aim of this audit was to assess which are the most effective and appropriate methods of reviewing quality of care from medical records. Further work is currently being undertaken to test out the conclusions and to assess whether it is possible to demonstrate a linkage between quality of care and outcomes.

All possible safeguards to preserve the quality of the data collected have been made by the University of Sheffield. Nevertheless it is important to interpret your results in this report using your knowledge of your own service and any difficulties you experienced in collecting your audit data that may have affected your own outcomes.

We are grateful to everyone who has helped with the project and appreciate the very considerable amount of time and effort that has gone into obtaining local data. We very much hope that this information will be useful for local audit purposes.

**Introduction**

Nine hospitals took part in this first phase of the audit. 1484 textual record reviews and 1400 criterion based record reviews were returned to the study team. Records were reviewed from two specialties; COPD and Heart Failure. Reviewers were either Nurses, Non-clinical audit staff, SpRs or other clinical staff.

This report presents the results of the audit of COPD records undertaken by reviewer 5732. Textual reviews were undertaken on 38 out of 50 records and criterion based reviews were undertaken on 36 out of the same 50 records. Reviews are of patients admitted with an exacerbation of COPD and who had a primary diagnosis of COPD during the time period 1st September 2004 - 28th February 2005.

**Methods**

* Textual review: quality of care is assessed using the reviewer’s own professional opinion
* Criterion based review: quality of care is assessed using a set of specific criteria

A phases of care approach was adopted for both review methods. For textual review, reviewers were asked to comment on the care received by the patient in the admission, initial management and pre-discharge phases. They were then asked to make a final overall comment. Reviewers were also asked to rate each the quality of care in phase on a 6 point scale and to rate the quality of the overall care on a 10 point scale.

For criterion based review, criteria were grouped under the phase of care headings used in the holistic review. Reviewers were asked to answer the questions using information from the patient record.

**Training for data collectors**

Two training days were held for data collectors. The training days were both held in London (Royal College of Pathologists and BMA House).

The training provided an introduction to the review methods and familiarised the reviewers with the materials to be used whilst undertaking the reviews (data collection software and review help notes). There were also sessions on recognising care quality variance when using the two review methods. For these sessions, examples were used from stroke care. This was so that we did not influence reviewers’ perceptions of care quality for the two audit conditions. Where reviewers were unable to attend a training day, 2 of the project team visited their hospital and provided training onsite. Reviewers who were unable to be present during the site visit were trained via telephone. Where possible, these reviewers were also assigned a “buddy” (someone at their hospital who had undertaken the training in a face to face setting).

* 15 reviewers were trained at the training days
* 10 reviewers were trained during a site visit
* 14 reviewers were provided with telephone training

The project team were available throughout the data collection period to answer queries and provide support and advice.

**Participants**

9 Hospitals took part in the COPD audit and 8 of those hospitals also took part in the Heart Failure audit. Hospitals were randomly selected to participate in the audit and consultants at each hospital were approached to provide their approval for the audit to take place and, to assist in finding staff to review records.

The COPD audit involved the following reviewers:

|  |  |
| --- | --- |
| **Staff type** | **Number of reviewers** |
| Doctor | 6 |
| Nurse | 6 |
| Non-clinical audit staff | 6 |
| Clinical other (e.g. physio or pharmacist) | 2 |
| **Total** | **20 reviewers** |

The Heart failure audit involved the following reviewers:

|  |  |
| --- | --- |
| **Staff type** | **Number of reviewers** |
| Doctor | 10 |
| Nurse | 5 |
| Non-clinical audit staff | 3 |
| Clinical other (e.g. physio or pharmacist) | 1 |
| **Total** | **19 reviewers** |

**Data return**

Reviewers were asked to review 50 Heart failure or COPD records, using each of the review methods (resulting in 100 record reviews). If all reviewers had returned all reviews, this would have resulted in a total of 3900 reviews.

However, not all reviewers were able to return the full amount of reviews. This was for a variety of reasons, for example staff changing jobs. This was particularly a problem for the SpR reviewers, some of whom rotated to a post in a different hospital during the audit period. Also, there were some difficulties in recruiting reviewers in some hospitals. This meant that these reviewers started the audit later than other reviewers and, as such, had less time to complete all the reviews. Due to work pressures, one hospital site was unable to return any Heart Failure reviews.

**Percentage of data returned**

|  |  |  |  |
| --- | --- | --- | --- |
| **Condition** | **Review type** | **Total number of reviews returned** | **% of reviews returned** |
| COPD | Textual | 901 | 90% |
| COPD | Criterion based | 834 | 83% |
| Heart Failure | Textual | 581 | 61% |
| Heart Failure | Criterion based | 563 | 59% |
| **Total** |  | **2879** | **74%** |

**Textual data**

We are using the textual data from the holistic review to investigate whether different staff types (for example audit staff, nurses and SpRs) make different types of comment or comment on different issues when asked to review quality of care from patient records.

Each reviewer’s comments have been coded according to the type of comment i.e. whether the comment is a judgement of the care provided or a description of the care provided. Comments have also been coded according to whether the comment relates to the patient records or patient care. We will use this information to determine which type of reviewer (SpR, Nurse, Non-clinical audit staff) provide the most useful types of comments about quality of care.

In some hospitals, different types of staff have reviewed the same records. This is so that we can compare the comments to gain an understanding of the types of comments made by different staff. All the analysis has been anonymised and each reviewer is only identifiable by their reviewer ID.

This analysis is ongoing. We hope to publish the results of the study findings and will provide you with details of any publications.

**Results**

**The following results relate to a review of 38 holistic and 36 criterion based patient records by reviewer 5732**

**Key data**

**Gender**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | Female | 18 | 50.0 |
|  | Male | 18 | 50.0 |
|  | Total | 36 | 100.0 |

**Number of patients accepted onto an early discharge scheme**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | Not accepted onto an early discharge scheme | 36 | 100.0 |
|  | not applicable | 0 | 0.0 |
|  | Accepted onto an early discharge scheme | 0 | 0.0 |
|  | Total | 0 | 0.0 |

**Number of patients discharged or died**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | Discharge | 35 | 97.2 |
|  | Died from COPD or complications of COPD | 0 | 0 |
|  | Not recorded | 1 | 2.8 |
|  | Total | 36 | 100.0 |

**Number of patients with previous admissions for COPD**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | No previous admissions for COPD | 19 | 52.8 |
|  | Previous admissions for COPD | 17 | 47.2 |
|  | Total | 36 | 100.0 |

**History and Patient Characteristics**

**Smoking status**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | current smoker | 15 | 41.7 |
|  | ex-smoker(stopped more than 3 months) | 18 | 50.0 |
|  | life long non-smoker | 2 | 5.6 |
|  | not recorded | 1 | 2.8 |
|  | Total | 36 | 100.0 |

**Social circumstances**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | lives alone with social service support | 1 | 2.8 |
|  | lives alone, no support | 11 | 30.6 |
|  | lives in nursing/residential home | 3 | 8.3 |
|  | lives in warden controlled (sheltered) housing | 1 | 2.8 |
|  | lives with spouse, close relative or carer | 20 | 55.6 |
|  | Total | 36 | 100.0 |

**Admission Phase**

**Please note, some of these results are subjective and are the opinions of the individual reviewers**

**Rating scale results: quality of care ratings for the admission phase**



**1 = unsatisfactory**

**6 = very best care**

**Quality of care rating for the admission phase: all reviewers from site 439**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reviewer** | **Mean quality of care rating** | **Standard deviation** | **Median** | **Range** |
| 5731 | 4.2 | 1.08 | 5.0 | 1.0 – 6.0 |
| 5732 | 4.4 | 0.72 | 4.0 | 3.0 – 6.0 |
| 5833 | 4.1 | 1.26 | 5.0 | 1.0 – 6.0 |
| 5834 | 5.0 | 0.99 | 5.0 | 2.0 – 6.0 |

**Quality of care rating for the admission phase: all hospitals**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hospital** | **Mean** | **Standard deviation** | **Median** | **Range** |
| Hospital 203 | 4.8 | 0.7 | 5.0 | 2.0 - 6.0 |
| Hospital 211 | 4.9 | 0.6 | 5.0 | 1.0 - 6.0 |
| Hospital 260 | 4.3 | 0.6 | 4.0 | 3.0 - 5.0 |
| Hospital 271 | 4.5 | 1.2 | 5.0 | 1.0 - 6.0 |
| Hospital 415 | 4.1 | 1.0 | 4.0 | 1.0 - 6.0 |
| Hospital 420 | 4.0 | 1.4 | 4.0 | 2.0 - 6.0 |
| Hospital 439 | 4.5 | 1.1 | 5.0 | 1.0 - 6.0 |
| Hospital 441 | 5.4 | 1.0 | 6.0 | 2.0 - 6.0 |
| Hospital 452 | 4.0 | 1.4 | 4.0 | 1.0 - 6.0 |

**Criterion based review: admission phase**

**Level of breathlessness**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | increased | 36 | 100.0 |
|  | Not increased | 0 | 0 |
|  | Total | 36 | 100.0 |

**Level of sputum**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | increased | 3 | 8.3 |
|  | not increased | 7 | 19.4 |
|  | not recorded | 26 | 72.2 |
|  | Total | 36 | 100.0 |

**Changes in sputum colour**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | changed | 3 | 8.3 |
|  | not changed | 5 | 13.9 |
|  | not recorded | 28 | 77.8 |
|  | Total | 36 | 100.0 |

**Sputum colour**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no sputum | 9 | 25.0 |
|  | not recorded | 6 | 16.7 |
|  | white or grey | 10 | 27.8 |
|  | yellow or green | 11 | 30.6 |
|  | Total | 36 | 100.0 |

**Was the dyspnoea rating recorded?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 12 | 33.3 |
|  | yes | 24 | 66.7 |
|  | Total | 36 | 100.0 |

**Performance status**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | limited activity but self care | 7 | 19.4 |
|  | limited self care | 4 | 11.1 |
|  | normal activity | 9 | 25.0 |
|  | not known | 13 | 36.1 |
|  | strenuous activity limited | 3 | 8.3 |
|  | Total | 36 | 100.0 |

**Chest x-ray within 24h?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 2 | 5.6 |
|  | yes | 34 | 94.4 |
|  | Total | 36 | 100.0 |

**If yes, is x-ray report in notes?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | Missing | 2 | 5.6 |
|  | No | 0 | 0.0 |
|  | yes | 34 | 94.4 |
|  | Total | 36 | 100.0 |

**Respiratory rate within 24h?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | yes | 2 | 5.6 |
|  | No | 34 | 94.4 |
|  | Total | 36 | 100.0 |

**Blood gases within 24h?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 8 | 22.2 |
|  | yes | 28 | 77.8 |
|  | Total | 36 | 100.0 |

**If yes, first reading**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  |  | 2 | 5.6 |
|  | 13 | 1 | 2.8 |
|  | 14 | 1 | 2.8 |
|  | 15 | 2 | 5.6 |
|  | 16 | 2 | 5.6 |
|  | 19 | 2 | 5.6 |
|  | 20 | 4 | 11.1 |
|  | 22 | 1 | 2.8 |
|  | 23 | 1 | 2.8 |
|  | 24 | 5 | 13.9 |
|  | 26 | 2 | 5.6 |
|  | 28 | 4 | 11.1 |
|  | 30 | 3 | 8.3 |
|  | 32 | 1 | 2.8 |
|  | 34 | 1 | 2.8 |
|  | 36 | 1 | 2.8 |
|  | 40 | 3 | 8.3 |
|  | Total | 36 | 100.0 |

**ECG performed?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 8 | 22.2 |
|  | yes | 28 | 77.8 |
|  | Total | 36 | 100.0 |

**Urea recorded?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 3 | 8.3 |
|  | yes | 33 | 91.7 |
|  | Total | 36 | 100.0 |

**Serum albumin recorded?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 8 | 22.2 |
|  | yes | 28 | 77.8 |
|  | Total | 36 | 100.0 |

**Record of medications at admission?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 3 | 8.3 |
|  | yes | 33 | 91.7 |
|  | Total | 36 | 100.0 |

**If peripheral oedema present, was it**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | No peripheral oedema/ not recorded | 20 | 55.6 |
|  | leg/ankles | 15 | 41.7 |
| sacral | 0 | 0.0 |
|  | not recorded | 1 | 2.8 |
|  | Total | 36 | 100.0 |

**Temperature at admission**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  |  | 1 | 2.8 |
|  | 35 | 1 | 2.8 |
|  | 35.4 | 1 | 2.8 |
|  | 35.5 | 1 | 2.8 |
|  | 35.7 | 1 | 2.8 |
|  | 36 | 1 | 2.8 |
|  | 36.1 | 3 | 8.3 |
|  | 36.2 | 2 | 5.6 |
|  | 36.3 | 2 | 5.6 |
|  | 36.4 | 3 | 8.3 |
|  | 36.5 | 2 | 5.6 |
|  | 36.7 | 3 | 8.3 |
|  | 36.8 | 1 | 2.8 |
|  | 36.9 | 2 | 5.6 |
|  | 37 | 4 | 11.1 |
|  | 37.1 | 1 | 2.8 |
|  | 37.2 | 2 | 5.6 |
|  | 37.4 | 1 | 2.8 |
|  | 37.5 | 1 | 2.8 |
|  | 37.8 | 2 | 5.6 |
|  | 38.2 | 1 | 2.8 |
|  | Total | 36 | 100.0 |

**Spirometry reading this admission?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 24 | 66.7 |
|  | yes | 12 | 33.3 |
|  | Total | 36 | 100.0 |

**Record of peripheral oedema?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | not recorded | 5 | 13.9 |
|  | yes - not present | 15 | 41.7 |
|  | yes - present | 16 | 44.4 |
|  | Total | 36 | 100.0 |

**Initial Management phase**

**Rating scale results: quality of care ratings for the initial management phase**

**Please note, some of these results are subjective and are the opinions of the individual reviewers**



**1 = unsatisfactory**

**6 = very best care**

**Quality of care rating for the initial management phase: all reviewers from site 439**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reviewer** | **Mean quality of care rating** | **Standard deviation** | **Median** | **Range** |
| 5731 | 4.5 | 0.92 | 5.0 | 2.0 – 6.0 |
| 5732 | 4.4 | 0.88 | 5.0 | 2.0 – 6.0 |
| 5833 | 4.4 | 1.33 | 5.0 | 1.0 – 6.0 |
| 5834 | 5.2 | 0.83 | 5.0 | 3.0 – 6.0 |

**Quality of care rating for the initial management phase: all hospitals**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hospital** | **Mean** | **Standard deviation** | **Median** | **Range** |
| Hospital 203 | 4.9 | 0.8 | 5.0 | 2.0 - 6.0 |
| Hospital 211 | 4.9 | 0.8 | 5.0 | 1.0 - 6.0 |
| Hospital 260 | 4.5 | 0.5 | 4.0 | 3.0 - 5.0 |
| Hospital 271 | 4.6 | 1.1 | 5.0 | 1.0 - 6.0 |
| Hospital 415 | 4.2 | 1.0 | 4.0 | 1.0 - 6.0 |
| Hospital 420 | 3.8 | 1.5 | 3.0 | 1.0 - 6.0 |
| Hospital 439 | 4.6 | 1.1 | 5.0 | 1.0 - 6.0 |
| Hospital 441 | 5.3 | 0.9 | 6.0 | 2.0 - 6.0 |
| Hospital 452 | 4.0 | 1.3 | 4.0 | 1.0 -6.0 |

**Criterion based review: initial management phase**

**Were antibiotics prescribed?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 5 | 13.9 |
|  | yes | 31 | 86.1 |
|  | Total | 36 | 100.0 |

**Were nebulised bronchodilators prescribed?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 3 | 8.3 |
|  | yes | 33 | 91.7 |
|  | Total | 36 | 100.0 |

**Did patient receive systemic corticosteroids?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 2 | 5.6 |
|  | yes | 34 | 94.4 |
|  | Total | 36 | 100.0 |

**Number of arterial blood gas results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | 0 | 7 | 19.4 |
|  | 1 | 18 | 50.0 |
|  | 2 | 2 | 5.6 |
|  | 3 | 4 | 11.1 |
|  | 4 | 2 | 5.6 |
|  | 5 | 1 | 2.8 |
|  | 6 | 2 | 5.6 |
|  | Total | 36 | 100.0 |

**pH less than 735 at any time?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | Missing data | 6 | 16.7 |
|  | no | 17 | 47.2 |
|  | yes | 13 | 36.1 |
|  | Total | 36 | 100.0 |

**Pre-discharge phase**

**Rating scale results: quality of care ratings for the pre-discharge phase**

**Please note, some of these results are subjective and are the opinions of the individual reviewers**



**1 = unsatisfactory**

**6 = very best care**

**Quality of care rating for the pre-discharge phase: all reviewers from site 439**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reviewer** | **Mean quality of care rating** | **Standard deviation** | **Median** | **Range** |
| 5731 | 4.5 | 0.85 | 5.0 | 2.0 – 5.0 |
| 5732 | 4.4 | 0.89 | 5.0 | 2.0 – 6.0 |
| 5833 | 4.4 | 1.14 | 5.0 | 1.0 – 6.0 |
| 5834 | 4.6 | 1.23 | 5.0 | 1.0 – 6.0 |

**Quality of care rating for the pre-discharge phase: all hospitals**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hospital** | **Mean** | **Standard deviation** | **Median** | **Range** |
| Hospital 203 | 4.74 | 0.76 | 5.0 | 1 - 6 |
| Hospital 211 | 4.8 | 0.84 | 5.0 | 1 - 5 |
| Hospital 260 | 4.3 | 0.57 | 4.0 | 3 - 5 |
| Hospital 271 | 4.5 | 1.36 | 5.0 | 1 - 6 |
| Hospital 415 | 4.2 | 1.16 | 4.0 | 1 -6 |
| Hospital 420 | 3.4 | 1.67 | 3.0 | 1 -6 |
| Hospital 439 | 4.5 | 1.14 | 5.0 | 1 - 6 |
| Hospital 441 | 4.8 | 0.98 | 5.0 | 2 -6 |
| Hospital 452 | 3.6 | 1.53 | 3.0 | 1 - 6 |

**Criterion based audit: pre-discharge phase**

**Oximetry within 48h of discharge?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | yes | 13 | 36.1 |
|  | no | 23 | 63.9 |
|  | Total | 36 | 100.0 |

**Assessment of home circumstances?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 30 | 83.3 |
|  | yes | 6 | 16.7 |
|  | Total | 36 | 100.0 |

**Where was patient discharged to?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | not applicable - died in hospital | 1 | 2.8 |
|  | nursing or residential care | 1 | 2.8 |
|  | other hospital | 1 | 2.8 |
|  | own home - independent of help | 25 | 69.4 |
|  | own home - with additional social support | 4 | 11.1 |
|  | sheltered housing or living with relative | 4 | 11.1 |
|  | Total | 36 | 100.0 |

**Discharge letter to primary care team?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | no | 4 | 11.1 |
|  | yes | 32 | 88.9 |
|  | Total | 36 | 100.0 |

**If yes, is there a clear list of medications?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | No discharge letter | 4 | 11.1 |
|  | no | 1 | 2.8 |
|  | yes | 31 | 86.1 |
|  | Total | 36 | 100.0 |

**Type of consultant at discharge?**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | n | Percent |
|  | care of elderly physician | 5 | 13.9 |
|  | general physician | 9 | 25.0 |
|  | other | 2 | 5.6 |
|  | respiratory physician | 20 | 55.6 |
|  | Total | 36 | 100.0 |

**Overall care**

**Rating scale results: quality of care ratings for the overall care**

**Please note, some of these results are subjective and are the opinions of the individual reviewers**



1 = unsatisfactory

10 = very best care

**Quality of care rating overall: all reviewers from site 439**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reviewer** | **Mean quality of care rating** | **Standard deviation** | **Median** | **Range** |
| 5731 | 7.4 | 1.90 | 8.0 | 2.0 – 10.0 |
| 5732 | 7.3 | 1.31 | 7.0 | 3.0 – 9.0 |
| 5833 | 7.4 | 2.21 | 8.0 | 2.0 – 10.0 |
| 5834 | 7.9 | 1.54 | 8.0 | 3.0 – 10.0 |

**Overall quality of care rating: all hospitals**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hospital** | **Mean** | **Standard deviation** | **Median** | **Range** |
| Hospital 203 | 8.0 | 1.26 | 8.0 | 3 - 10 |
| Hospital 211 | 8.2 | 1.46 | 9.0 | 2 - 9 |
| Hospital 260 | 7.2 | 1.00 | 7.0 | 4 -9 |
| Hospital 271 | 7.6 | 1.90 | 8.0 | 1 -10 |
| Hospital 415 | 7.3 | 1.50 | 8.0 | 1 -10 |
| Hospital 420 | 5.9 | 2.40 | 6.0 | 2 -10 |
| Hospital 439 | 7.5 | 1.79 | 8.0 | 2 -10 |
| Hospital 441 | 8.3 | 1.25 | 8.0 | 6 -10 |
| Hospital 452 | 5.7 | 2.09 | 6.0 | 1 -10 |

**Patient records**

**Rating scale results: ratings for the quality of patient records**



**1 = inadequate**

**6 = excellentAppendix 7**

**COPD – correlations between holistic mean overall scale scores (1 = unsatisfactory to 6 = very best care) and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of COPD patients who die in hospital within 28 days | -0.295 | 0.207 | Weak |
| HSMR from Dr. Foster (3 year mortality) | -0.157 | 0.508 | Very weak |
| HSMR from Dr. Foster (1 year mortality) | -0.072 | 0.763 | Very weak |
| Incidents to NPSA per 100 bed days | -0.345 | 0.136 | Weak |
| SMR for deaths in low mortality HRG’s | 0.049 | 0.838 | Very weak |
| COPD finished consultant episodes | 0.125 | 0.610 | Very weak |
| COPD bed days | 0.102 | 0.677 | Very weak |
| COPD mean length of stay | -0.167 | 0.495 | Very weak |
| COPD mean age | 0.011 | 0.963 | Very weak |
| Star rating (0 worst to 3 best) | -0.077\* | 0.746 | Very weak |
| Use of resources HCC | -0.184\* | 0.437 | Very weak |
| Patient’s experience | 0.222 | 0.347 | Weak |
| Quality of services | -0.118\* | 0.620 | Very weak |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.185 | 0.448 | Very weak |
| Existing national targets | -0.072\* | 0.764 | Very weak |
| New national targets | 0.105\* | 0.658 | Very weak |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | -0.328 | 0.158 | Weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | -0.490 | 0.028 | Moderate |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | -0.131 | 0.581 | Very weak |
| NHS staff survey Q24a: Know how to report (% yes) | -0.230 | 0.329 | Weak |
| NHS staff survey Q24b: System for reporting (% yes) | -0.243 | 0.303 | Weak |
| NHS staff survey Q24b: System for reporting (% no) | -0.154 | 0.516 | Very weak |
| NHS staff survey Q24b: System for reporting (% DK) | 0.294 | 0.208 | Very weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | -0.503 | 0.024 | Moderate |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | -0.282 | 0.228 | Weak |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 8**

**Heart Failure – correlations between holistic mean overall scale scores (1 = unsatisfactory to 6 = very best care) and outcome variables**

*Correlations*

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of HF patients who die in hospital within 28 days | -0.334 | 0.149 | Weak |
| HSMR from Dr. Foster (3 year mortality) | -0.183 | 0.439 | Very weak |
| HSMR from Dr. Foster (1 year mortality) | -0.308 | 0.186 | Weak |
| Incidents to NPSA per 100 bed days | -0.228 | 0.335 | Weak |
| SMR for deaths in low mortality HRG’s | -0.093 | 0.696 | Very weak |
| HF finished consultant episodes | -0.174 | 0.477 | Very weak |
| HF bed days | -0.237 | 0.329 | Weak |
| HF mean length of stay | 0.064 | 0.795 | Very weak |
| HF mean age | -0.445 | 0.056 | Moderate |
| Star rating (0 worst to 3 best) | 0.240\* | 0.309 | Weak |
| Use of resources HCC | 0.345\* | 0.136 | Weak |
| Patient’s experience | -0.365 | 0.114 | Weak |
| Quality of services | 0.651\* | 0.002 | Strong |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.350 | 0.142 | Weak |
| Existing national targets | 0.765\* | < 0.001 | Strong |
| New national targets | 0.453\* | 0.045 | Moderate |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | -0.261 | 0.267 | Weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | 0.308 | 0.187 | Weak |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | 0.430 | 0.059 | Moderate |
| NHS staff survey Q24a: Know how to report (% yes) | 0.509 | 0.022 | Moderate |
| NHS staff survey Q24b: System for reporting (% yes) | 0.264 | 0.261 | Weak |
| NHS staff survey Q24b: System for reporting (% no) | 0.126 | 0.598 | Very weak |
| NHS staff survey Q24b: System for reporting (% DK) | -0.306 | 0.189 | Weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | 0.442 | 0.051 | Moderate |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | 0.078 | 0.744 | Very weak |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 9**

**COPD – correlations between holistic mean phase scale scores and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of COPD patients who die in hospital within 28 days | -0.290 | 0.215 | Weak |
| HSMR from Dr. Foster (3 year mortality) | -0.135 | 0.569 | Very weak |
| HSMR from Dr. Foster (1 year mortality) | -0.095 | 0.691 | Very weak |
| Incidents to NPSA per 100 bed days | -0.067 | 0.778 | Very weak |
| SMR for deaths in low mortality HRG’s | 0.070 | 0.771 | Very weak |
| COPD finished consultant episodes | 0.422 | 0.072 | Moderate |
| COPD bed days | 0.387 | 0.102 | Weak |
| COPD mean length of stay | -0.146 | 0.550 | Very weak |
| COPD mean age | 0.108 | 0.658 | Very weak |
| Star rating (0 worst to 3 best) | 0.035\* | 0.882 | Very weak |
| Use of resources HCC | -0.054\* | 0.821 | Very weak |
| Patient’s experience | -0.365 | 0.114 | Weak |
| Quality of services | -0.083 | 0.729 | Very weak |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.350 | 0.142 | Weak |
| Existing national targets | -0.054\* | 0.821 | Very weak |
| New national targets | 0.073\* | 0.759 | Very weak |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | -0.236 | 0.316 | Weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | -0.330 | 0.156 | Weak |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | 0.038 | 0.872 | Very weak |
| NHS staff survey Q24a: Know how to report (% yes) | -0.248 | 0.292 | Weak |
| NHS staff survey Q24b: System for reporting (% yes) | -0.203 | 0.390 | Weak |
| NHS staff survey Q24b: System for reporting (% no) | -0.103 | 0.664 | Very weak |
| NHS staff survey Q24b: System for reporting (% DK) | 0.238 | 0.312 | Weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | -0.454 | 0.044 | Moderate |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | -0.268 | 0.252 | Weak |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 10**

**Heart Failure – correlations between holistic mean overall scale scores and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of HF patients who die in hospital within 28 days | -0.274 | 0.242 | Weak |
| HSMR from Dr. Foster (3 year mortality) | -0.139 | 0.560 | Very weak |
| HSMR from Dr. Foster (1 year mortality) | -0.272 | 0.245 | Weak |
| Incidents to NPSA per 100 bed days | -0.215 | 0.362 | Weak |
| SMR for deaths in low mortality HRG’s | -0.064 | 0.789 | Very weak |
| HF finished consultant episodes | -0.147 | 0.549 | Very weak |
| HF bed days | -0.226 | 0.353 | Weak |
| HF mean length of stay | 0.029 | 0.905 | Very weak |
| HF mean age | -0.552 | 0.014 | Moderate |
| Star rating (0 worst to 3 best) | 0.222\* | 0.347 | Weak |
| Use of resources HCC | 0.248\* | 0.292 | Weak |
| Patient’s experience | -0.292 | 0.211 | Weak |
| Quality of services | 0.486\* | 0.030 | Moderate |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.463 | 0.046 | Moderate |
| Existing national targets | 0.691\* | 0.001 | Strong |
| New national targets | 0.226\* | 0.338 | Weak |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | -0.212 | 0.369 | Weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | 0.331 | 0.154 | Weak |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | 0.470 | 0.037 | Moderate |
| NHS staff survey Q24a: Know how to report (% yes) | 0.546 | 0.013 | Moderate |
| NHS staff survey Q24b: System for reporting (% yes) | 0.286 | 0.222 | Weak |
| NHS staff survey Q24b: System for reporting (% no) | 0.076 | 0.750 | Very weak |
| NHS staff survey Q24b: System for reporting (% DK) | -0.313 | 0.180 | Weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | 0.470 | 0.037 | Moderate |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | 0.009 | 0.972 | Very weak |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 11**

**COPD – correlations between holistic mean criterion scores and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of COPD patients who die in hospital within 28 days | -0.297 | 0.203 | Weak |
| HSMR from Dr. Foster (3 year mortality) | -0.247 | 0.295 | Weak |
| HSMR from Dr. Foster (1 year mortality) | -0.111 | 0.640 | Very weak |
| Incidents to NPSA per 100 bed days | -0.489 | 0.029 | Moderate |
| SMR for deaths in low mortality HRG’s | 0.308 | 0.187 | Weak |
| COPD finished consultant episodes | -0.022 | 0.928 | Very weak |
| COPD bed days | -0.007 | 0.979 | Very weak |
| COPD mean length of stay | 0.118 | 0.629 | Very weak |
| COPD mean age | 0.125 | 0.611 | Very weak |
| Star rating (0 worst to 3 best) | 0.019\* | 0.936 | Very weak |
| Use of resources HCC | -0.311\* | 0.182 | Weak |
| Patient’s experience | 0.101 | 0.672 | Very weak |
| Quality of services | -0.188\* | 0.427 | Very weak |
| Percentage of patient’s with acute MI receiving thrombolysis | 0.042 | 0.866 | Very weak |
| Existing national targets | -0.207\* | 0.381 | Weak |
| New national targets | -0.049\* | 0.838 | Very weak |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | 0.069 | 0.772 | Very weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | -0.238 | 0.312 | Weak |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | -0.084 | 0.726 | Very weak |
| NHS staff survey Q24a: Know how to report (% yes) | -0.263 | 0.263 | Very weak |
| NHS staff survey Q24b: System for reporting (% yes) | 0.093 | 0.698 | Very weak |
| NHS staff survey Q24b: System for reporting (% no) | -0.131 | 0.582 | Very weak |
| NHS staff survey Q24b: System for reporting (% DK) | -0.052 | 0.827 | Very weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | -0.139 | 0.558 | Very weak |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | 0.475 | 0.034 | Moderate |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 12**

**Heart Failure – correlations between mean criterion scores and outcome variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Correlation coefficient | p-value | Relationship |
| Percentage of HF patients who die in hospital within 28 days | -0.357 | 0.122 | Weak |
| HSMR from Dr. Foster (3 year mortality) | 0.014 | 0.952 | Very weak |
| HSMR from Dr. Foster (1 year mortality) | -0.218 | 0.357 | Weak |
| Incidents to NPSA per 100 bed days | 0.528 | 0.017 | Moderate |
| SMR for deaths in low mortality HRG’s | 0.280 | 0.232 | Weak |
| HF finished consultant episodes | -0.255 | 0.291 | Weak |
| HF bed days | -0.259 | 0.284 | Weak |
| HF mean length of stay | 0.017 | 0.944 | Very weak |
| HF mean age | -0.325 | 0.172 | Weak |
| Star rating (0 worst to 3 best) | -0.254\* | 0.174 | Weak |
| Use of resources HCC | 0.029\* | 0.279 | Very weak |
| Patient’s experience | -0.377 | 0.903 | Weak |
| Quality of services | -0.033\* | 0.891 | Very weak |
| Percentage of patient’s with acute MI receiving thrombolysis | -0.102 | 0.679 | Very weak |
| Existing national targets | -0.147\* | 0.535 | Very weak |
| New national targets | 0.155\* | 0.515 | Very weak |
| NHS staff survey Q25a: Seen errors in the past month (% yes) | 0.192 | 0.416 | Very weak |
| NHS staff survey Q27b: Encouraged to report errors (mean) | 0.137 | 0.564 | Very weak |
| NHS staff survey Q27e: Trust takes action to ensure does not happen again (mean) | -0.092 | 0.698 | Very weak |
| NHS staff survey Q24a: Know how to report (% yes) | -0.142 | 0.551 | Very weak |
| NHS staff survey Q24b: System for reporting (% yes) | -0.145 | 0.543 | Very weak |
| NHS staff survey Q24b: System for reporting (% no) | 0.620 | 0.004 | Strong |
| NHS staff survey Q24b: System for reporting (% DK) | -0.051 | 0.833 | Very weak |
| NHS staff survey Q22e: Care of patient/service user is top priority (mean) | 0.336 | 0.147 | Weak |
| NHS staff survey Q22f: Happy with standard of care provided (mean) | -0.291 | 0.214 | Weak |

\* Spearman’s rank correlation used

Correlations at the < 0.05 significance level highlighted

**Appendix 13**

**Comparison of holistic and criterion based review methods using structured clinical records in stroke care**

**Background**

UK stroke care tends to have structured medical records for hospital inpatients, with prospective completion of patient records based on structured phases of care, in some units. This, it is hypothesised, changes the type and quality of data collected in the medical record and thus may more accurately enable conformance with, and assessment of, good care standards.

It is currently unknown whether the use of structured medical records has any effect on the quality of information available for peer review, or whether it might differentially influence the quality of information that is captured by explicit or implicit review methods. It may also be that there is a higher level of inter-rater reliability to be found between and within types of reviewers when using structured, prospective record keeping, compared with that found in the main research project.

Major national sentinel audit projects in the UK have already used review criteria based, explicit review methods to explore quality variance in stroke care. [1] These were undertaken by teams of nurses or physicians trained in records review methods. Substantial variations in organisation and clinical care have been identified across the 8200 cases included in the national stroke audit. The audit did not use a holistic approach, which has been hypothesised as an alternative means of identifying quality variation, especially in complex cases. [2].

**Study questions**

This small adjunct study seeks to answer two related questions. First, what are the similarities and differences in peer review information captured by explicit (review criterion based) methods and implicit (holistic) methods from structured (stroke care) clinical records? Second, are there differences in the type of information recorded by clinical audit staff (including nurses) and by doctors, using the two types of review methods?

This study is nested within the main medical records study which addresses similar research questions but in which the records are not structured. Although the stroke care study is small, the overall results of the stroke care study can therefore be contrasted with the non-structured record results in the Phase One study to begin to explore whether there are differences between the type of information that can be extracted from the two different types of records.

*Secondary study questions:*

##### Does structured prospective medical record keeping in stroke care influence the type, extent and quality of data recorded in clinical audit review (as compared with the type and quality of data found in unstructured record keeping in the Phase One study)?

* When using explicit (review criteria based) clinical audit review methods, does the use of structured recording in stroke care change the proportions of recorded criteria compared with the proportions recorded in unstructured records for COPD and heart failure care?
* For both explicit (review criteria based) and implicit (holistic) clinical audit case note review methods, does reliability improve between and within reviewer types when structured recording is used in comparison to unstructured recording?

**Methods**

The overall research approach was to investigate the impact of structured prospective record keeping on the reliability and completeness of holistic and criterion based case note review methods. Quality of care was assessed using a combined holistic and algorithmic method, as used in the Phase Two outcomes study, by one nursing trained reviewer and two medical reviewers. The same case notes were reviewed by each reviewer.

## **Stroke**

Stroke is recognised as the third biggest cause of death in the UK. It is also the largest single cause of severe disability in older people. In excess of 110,000 people in England each year will suffer from a stroke, which incurs NHS costs of over £2.8 billion per year.[3] All hospitals that care for stroke patients were required to have a specialist stroke service by 2004, as set out by the National Service Framework for Older People [4]. To support our choice of Stroke for this adjunct study, we took into account the availability of an evidence based guideline produced by the Royal College of Physicians, together with the existence of the National Sentinel Stroke Audit [1]. The National Sentinel Stroke Audit Criteria provided the basis for developing a set of review criteria for safety and quality assessment for stroke management for this study (see also the section on criterion development for the Phase One study).

***Number of case notes for review***

Each reviewer was asked to review 40 stroke care records, as in the Phase One study.

***Selection and recruitment of hospitals***

Only one hospital participated in this small study. This hospital was chosen because of the study team’s close links with the stroke care staff, whose input was crucial to the development of the review criteria and the running of the study. A second hospital was also approached but, whilst they were keen to participate, they did not have staff available to take part.

***Numbers and types of review and reviewers***

Two doctors in training (specialist registrars) and one clinical audit nurse were recruited to review 40 records of patients admitted to hospital for care for an acute stroke.

***Holistic review data capture***

Holistic review data was collected using the same methods as in the Phase One study. Reviewers were asked to provide a textual comment on the quality of care and also to rate the quality of care on a six point rating scale (1= unsatisfactory, 6 = very best care). This was done for each of 3 phases of care (admission, initial management and pre-discharge) and for care overall. Care overall was rated on a 10 point rating scale (1 = unsatisfactory, 10 = very best care).

***Assessing the quality of recording in the case notes***

Evaluation of quality of care through case note review is critically dependent on the quality of recording in the case notes, together with that in associated data sources such as computerised pathology and radiology results. In order to assess the quality and completeness of the records under review, reviewers were asked to assess the quality of each record using a six point rating scale (1 = inadequate, 6 = excellent), as per the main study.

***Review criteria development for stroke care***

The basis for the development of the review criteria was an already established criterion based audit dataset within the National Sentinel Stroke audit, an organisational and clinical audit comprising 94 criterion. We developed a shorter version of the clinical component of the national audit dataset through discussions with stroke care staff, using the same approach as the Phase One study.

For the stroke study criterion based questions, pre-defined answer options were provided for all questions. Usually, the options were “Yes”, “No” and “Not recorded”. Reviewers were instructed to answer “Yes” if the care was provided or the result of a test was in the patient record and “No” if the care was not provided for a valid reason. Examples of valid reasons were provided for each question and included things such as if the patient died, or was unconscious or was receiving palliative care. Reviewers were instructed to answer “Not recorded“ if the information they were looking for was missing from the record. Where information is missing from the record it is presumed that that care was not provided.

***Developing data capture tools***

Data collection materials were developed in Microsoft Access© and were designed to be easily used by the reviewers. The software had a facility to easily export formatted data to the study team.

***Staff training***

Where possible, we intended for staff reviewing records for the stroke care study attended one of the training days for the main study, although only the nursing trained clinical reviewer was able to attend the training day. The two doctors were unable to attend due to clinical commitments, so one attended a one-to-one training session with the study project manager, whilst the other was trained via telephone by the project manager.

***Analysis methods***

Holistic scale score analysis

Summary statistics for the holistic quality of care ratings were calculated for each reviewer, for the phases of care and overall data. Box and whisker plots comparing the quality of care ratings for the phases of care and overall quality of care for each reviewer were also produced.

Measuring reliability between reviewer pairs

The reliability between the reviewers overall quality of care ratings was assessed by calculating Interclass Correlation Coefficients (ICCs) in SPSS. ICCs were calculated for each reviewer pair (Doctor 1 v Doctor 2, Doctor 1 v Clinical audit 1 and Doctor 2 v Clinical audit 1) as well as a combined ICC for all three reviewers.

Criterion based review

The score for each criterion was summed to create quality of care scores for each phase of care and overall. Summary statistics for the criterion based quality of care score were calculated, as were box and whisker plots comparing the three reviewer’s data.

Measuring reliability between reviewer pairs

ICCs were used to assess the inter-rater reliability of the criterion based quality of care scores. As with the holistic reliability analysis, ICCs were calculated for each reviewer pair and for all three reviewers.

***Ethics review***

The ethics review was the same as that for the main studies.

***Healthcare governance***

Although the main study had already received ethical approval further discussions were held with the Clinical Effectiveness Manager of the study hospital to determine whether this small adjunct study was research, clinical audit or service review [5]. Ethical principles were considered following the decision that this work would be undertaken as service review and conducted in line with local governance procedures.**Results**

***Holistic quality of care rating scale***

Completeness of data capture

Each of the Doctor reviewers reviewed 37 out of the 40 records and the Clinical reviewer reviewed 40 out of 40 records. Only the 37 reviews that were reviewed by each of the three reviewers were included in the analysis. The level of completeness of holistic review was assessed by calculating the amount of missing data for each of the quality of care rating scales. The amount of missing data for each phase for each reviewer is presented in Table 13.1 which shows that the completeness level of the holistic rating scale is high. One reviewer (1733, Clinical) had no missing data. The amount of missing data for this section of the review is small. The most missing data was recorded by Doctor 1 for the Initial Management Phase, however, 8.8% missing data equates to 3 instances of missing data.

**Table 13.1: Holistic review completeness of data collection**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Phase of care | Number of reviews | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer |
| Admission phase missing data | 37 | 0% | 0% | 2.7% |
| Initial Management phase missing data | 37 | 0% | 8.8% | 2.7% |
| Pre-discharge phase missing data | 37 | 0% | 2.7% | 0% |
| Overall missing data | 37 | 0% | 0% | 2.7% |

Quality of care

Table 13.2 presents the results of the quality of care analysis of stroke care using a holistic rating scale, where 1 = unsatisfactory care and 6 = very best care. On the whole the reviewer mean/median results were similar for each phase of care. Doctor 2 tends to rate the quality of care lower than the Clinical reviewer and Doctor 1, particularly in the pre-discharge section. The clinical reviewer only uses a limited section of the rating scale (between 3 and 5), whereas the other two reviewers (doctors) tend to use all of the available scale.

**Table 13.2: Holistic scale score results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phase of Care** |  | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer |
| **Admission phase quality of care rating** | Mean  (SD) | 4.5  (0.5) | 4.7  (0.9) | 4.4  (1..3) |
| Median (IQR) | 4.0  (4.0 – 5.0) | 5.0  (4.0 – 5.0) | 5.0  (4.0 – 5.0) |
| Min - max | 4.0 – 5.0 | 1.0 – 6.0 | 2.0 – 6.0 |
| **Initial Management phase quality of care rating** | Mean  (SD) | 4.5 (0.6) | 4.88  (0.81) | 4.3  (1.1) |
| Median (IQR) | 5.00  (4.00 – 5.00) | 5.00  (4.00 – 5.00) | 5.0  (3.3 – 5.0) |
| Min - max | 3.0 – 5.0 | 3.0 – 6.0 | 2.0 – 6.0) |
| **Pre-discharge phase quality of care rating** | Mean  (SD) | 4.5  (0.6) | 5.0  (0.7) | 3.8  (1.1) |
| Median (IQR) | 5.0  (4.0 – 5.0) | 5.0  (5.0 – 5.0) | 4.0  (3.0 – 5.0) |
| Min - max | 3.0 – 5.0 | 3.0 – 6.0 | 1.0 – 5.0 |
| **Overall phase quality of care rating** | Mean  (SD) | 4.4  (0.5) | 4.8  (0.9) | 4.0  (0.9) |
| Median (IQR) | 4.0  (4.0 – 5.0) | 5.0  (4.5 – 5.0) | 4.0  (3.0 – 5.0) |
| Min - max | 4.0– 5.0 | 1.0 – 6.0 | 2.0 – 5.0 |

Box and whisker plots of holistic scale score data

The box and whisker plots compare the median quality of care scale rating and the inter-quartile range, for each reviewer for each phase of care.

**Figure 13.1: Holistic phase quality of care rating**

|  |  |
| --- | --- |
| **Admission phase** | **Initial management phase** |
|  |  |
| **Pre-discharge phase** | **Total score** |
|  |  |

|  |  |
| --- | --- |
| Key |  |
| 1733 | Clinical reviewer |
| 1831 | Doctor 1 |
| 1832 | Doctor 2 |

Reliability between reviewers

Table 13.3 presents the results of the stroke care holistic review inter-rater reliability analysis and shows the level of agreement between reviewers for the overall quality of care ratings.

**Table 13.3: Inter-rater reliability of holistic overall quality of care ratings**

|  |  |  |  |
| --- | --- | --- | --- |
| Reviewer Pair | Phase of care | ICC | Significance |
| Doctor 1 versus Doctor 2 | Overall quality of care rating | 0.328 | 0.022 |
| Doctor 1 versus Clinical audit 1 | Overall quality of care rating | - 0.285 | 0.959 |
| Doctor 2 versus Clinical 1 | Overall quality of care rating | 0.047 | 0.389 |
| All staff (Doctor 1, Doctor 2, Clinical 1) | Overall quality of care rating | 0.077 | 0.210 |

Comparison with results from the main study

The pair of doctor reviewers achieves the highest reliability. This is in line with findings from the main study, which found that pairs of doctor reviewers achieved the highest reliability for holistic review. The Phase One reliability study also found that there was low reliability between different staff type reviewer pairs e.g. doctor and nurse for holistic review (see Table 7). This finding is supported by the analysis of the stroke data as the different staff type reviewer pairs achieve lower ICCs here.

The amount of comparison that can be undertaken between the stroke reviewer reliability analysis and the main study reviewer pair reliability analysis is limited, due to the small number of stroke reviewers taking part. The doctor reviewer pair ICC is similar to that of the COPD doctor reviewer pair in the Phase One reliability study (0.328 versus 0.33 respectively, Table 6). In this much smaller study on stroke care the different staff type reviewer pair reliability comparisons are generally much lower than those presented in Table 6 in the Phase One study.

***Criterion based review***

Completeness of data capture

Each reviewer completed 37 reviews (of the same patient records). Where a reviewer did not select one of the pre-defined answer options for each criterion, this was classed as missing data. The results in Table 13.4 show that there are very low missing data rates for all reviewers for the criterion based data collection.

**Table 13.4: criterion based data completeness rates**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer | Total |
| Total number of reviews | 37 | 37 | 37 | 111 |
| Total number of data items available (sum of admission, initial management and pre-discharge phases) | 1406 | 1406 | 1406 | 4218 |
| Total number of data items missing | 0 | 19 | 7 | 26 |
| Percentage of data missing from each reviewer | 0% | 0.01% | 0.004% | 0.006% |

Criterion based quality of care scores

Quality of care scores were assigned to the criterion in the admission, initial management and pre-discharge phases. The method used was similar to that in the Phase One study, whereby each time a criterion was met or done or the reviewer selected the not done for a valid reason option a score of 1 was given. If a reviewer selected not recorded, the review item did not receive a score as this option presumes the care was not provided. The mean and median quality of care scores are presented in table 13.5.

**Table 13.5: Criterion based quality of care scores**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phase of care** |  | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer |
| Admission Phase (out of 11) | Mean  (SD) | 9.2  (1.2) | 10.5  (0.6) | 7.8  (2.3) |
| Median  (IQR) | 9.0  (9.0 – 10.0) | 11.0  (10.0 – 11.0) | 8.0  (6.0 – 10.0) |
| Min – Max | 6.0 – 11.0 | 9.0 – 11.0 | 3.0 – 11.0 |
| Initial Management Phase (out of 6) | Mean  (SD) | 5.2  (0.9) | 5.8  (0.4) | 4.6  (1.2) |
| Median  (IQR) | 5.0  (5.0 – 6.0) | 6.0  (6.0 – 6.0) | 5.0  (4.0 – 5.0) |
| Min – Max | 2.0 – 6.0 | 5.0 – 6.0 | 1.0 – 6.0) |
| Pre-discharge phase (out of 16) | Mean  (SD) | 13.5  (1.3) | 15.2  (1.9) | 10.1  (2.8) |
| Median  (IQR) | 13.0  (13.0 – 14.5) | 16.0  (15.0 – 16.0) | 10.0  (8.0 – 12.0) |
| Min – Max | 11.0 – 16.0 | 5.0 – 16.0 | 4.0 – 16.0 |
| Total score (sum of all phases, max 33) | Mean  (SD) | 28.0  (2.3) | 31.2  (2.0) | 22.6  (4.5) |
| Median  (IQR) | 29.0  (27.0 – 29.0) | 32.0  (31.0 – 32.0) | 22.6  (20.8 – 25.9) |
| Min – Max | 22.0 – 33.0 | 22.0 – 33.0 | 11.4 – 33.0 |

Box and whisker plots of criterion based total quality of care scores

The box and whisker plots compare the median criterion based quality of care scores and the inter-quartile range, for each reviewer for each phase of care.

**Figure 13.2: Criterion based quality of care scores**

|  |  |
| --- | --- |
| **Admission phase** | **Initial management phase** |
|  |  |
| **Pre-discharge phase** | **Total score** |
|  |  |

|  |  |
| --- | --- |
| Key |  |
| 1733 | Clinical reviewer |
| 1831 | Doctor 1 |
| 1832 | Doctor 2 |

Reliability between reviewers

Table 13.6 presents the results of the stroke criterion based review reliability analysis. The quality of care scores assigned to the criterion based data was used to calculate ICCs for each reviewer pair.

**Table 13.6: Inter-rater reliability between criterion based review quality of care scores**

|  |  |  |  |
| --- | --- | --- | --- |
| Reviewer Pair |  | ICC | Significance |
| Doctor 1 versus Doctor 2 | Criterion based total quality of care score | 0.031 | 0.426 |
| Doctor 1 versus Clinical 1 | Criterion based total quality of care score | 0.126 | 0.226 |
| Doctor 2 versus Clinical 1 | Criterion based total quality of care score | 0.384 | 0.009 |
| All staff (Doctor 1, Doctor 2, Clinical 1) | Criterion based total quality of care score | 0.199 | 0.022 |

Comparison with results from the Phase One study

As with the holistic data, the amount of comparison that can be undertaken between the stroke reviewer reliability analysis and the main study reliability analysis is limited, due to the small number of stroke reviewers taking part.

On the whole the reliability results for the stroke data are much lower than that of the Phase One study where the reliability results were 0.88 (range 0.64 to 0.96) for the pairs of doctor reviewers (Table 9). The stroke review reliability for the doctor pairs is much less at only 0.031, suggesting that the reviewers are not completing the data collection form in the same way. However, reliability results for doctors in the Phase One study did vary quite widely between individual pairs

Time taken to complete reviews

Data presented in Table 13.7 show that each stroke care record review took approximately 1 hour.

**Table 13.7: Summary statistics for time taken to review records (minutes)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer |
| Mean | 53.78 | 70.14 | 66.57 |
| Median | 50.00 | 70.00 | 60.00 |
| Std. Deviation | 12.985 | 12.047 | 26.031 |
| Minimum | 40 | 45 | 30 |
| Maximum | 110 | 90 | 150 |

The length of hospital stay for stroke patients tends to be long and the records associated with the care are large. From our sample, the mean length of stay for each patient was 33.7 days, but this ranged from 2 days to 314 days. Also, the method used was a joint holistic and criterion based method, with scale scores, textual data and review of criteria, so the length of time taken to perform each review is probably not unreasonable.

Quality of records

The quality of the case notes reviewed for this study was rated on a rating scale (1 = poor, 6 = excellent). Case notes received similar quality ratings from the Clinical and Doctor 1 reviewer, while Doctor 2 tended assign a lower quality rating than the other two reviewers (Table 13.8).

**Table 13.8: Quality of record ratings**

|  |  |  |  |
| --- | --- | --- | --- |
| **Quality of records rating** | Clinical reviewer (Nurse | Doctor 1 reviewer | Doctor 2 reviewer |
| Mean | 4.38 | 4.70 | 3.33 |
| (SD) | (0.5) | (0.661) | (0.717) |
| Median (IQR) | 4.0  (4.0 – 5.0) | 5.00  (4.0 – 5.0) | 3.00  (3.0 – 4.0) |
| Min - max | 3 – 5 | 3 – 6 | 1 – 4 |

In the Phase One study the mean quality of case notes rating for COPD and Heart Failure were 4.3 (SD 1.2) and 4.7 (SD 0.9) respectively. The stroke care case notes received similar ratings.

**Conclusions**

The size of this adjunct study was limited by resources and subsequently by access, thus reducing its generalisability. For the criterion based component of the review there is some indication that the reviewers were able to capture a more complete data set than the 39 reviewers in the main study were able to do. This may have been due to the quality of the recording and perhaps the structured nature of the case notes, although it is also possible that the reviewers were more skilled at the task than were the main study reviewers.

The inter-rater reliability results were poor for the holistic reviews, more so that the main study although even the main study showed that there were considerable differences between reviewers in their inter and intra rater reliability. There may be a number of reasons for this poor level of agreement, including the general level of difficulty of providing holistic reviews of case notes belonging to patients who had prolonged hospital stays. Under such conditions, it may be that holistic reviewing requires a very high level of training and experience to be able to identify variations in care from the mass of available information, perhaps supported by electronic means of screening such as might be possible by using trigger methodologies (Appendix 14) based on a condition specific set of review criteria.

**References**

1] Intercollegiate Stroke Audit Working Party. *National Sentinel Stroke Audit*. London: Royal College of Physicians of London, 2007

2] Mohammed MA, Mant J, Bentham L, Stevens A, Hussain S. Process and mortality of stroke patients with and without do not resuscitate order in the West Midlands, UK. *Int J Qual Healthcare*, 2006;**18**:102-106.

3] <http://www.nao.org.uk/pn/05-06/0506452.htm>

4] Department of Health. *National Service Framework for Older People*. London: Department of Health, 2001

5] Mawson S, Gerrish K, Schofield J, Debbage S, Somers A. A pragmatic governance framework for differentiating between research, audit and service review activities. *Clin in Management*, 2007;**15**:29-35

**Appendix 14**

**The place of trigger tool methodology in case note review for quality and safety**

**Context of the analysis**

The initial study proposal indicated that it would be valuable to undertake an exploration of electronic trigger tool methods for assessing safety and quality in the two study conditions, acute exacerbations of COPD and heart failure, in contrasting trigger tools with paper-based holistic and criterion review.

Because this trigger tool study required research ethics approval it could only practically be undertaken in a hospital local to the study team, given the extent of the resource commitments of the main studies. It proved to be the case that none of the local hospitals had sufficient electronic records systems to support even a small study. In the most likely setting for the research there were a total of seven separate paper based record systems. The research commissioners therefore agreed that a short review of trigger tool methods in the context of paper based records would be an appropriate alternative.

**Trigger tool methods**

Healthcare trigger tools were first described by Classen and colleagues[1] as an electronic screening tool for identifying markers or ‘sentinels’[2] for possible adverse drug events (ADEs). Since this original prototype was developed in the context of a hospital with an electronic record system, it was possible to develop data searching techniques that scanned for a drug, test or procedure that is usually associated with the management of an ADE. If a marker was found then a full review of a medical record could be undertaken to determine whether there had indeed been an avoidable event.

It should be noted here, however, that the term ADE includes both preventable adverse events and adverse drug reactions which may be unforeseeable even under circumstances of the very best care. So the sensitivity of the original electronic trigger tool system for identifying safety events was limited by the choice of tracer – in this case an ADE.

Resar et al[2] point out that this initial trigger tool system had the benefit of much reducing the staff time that might otherwise be needed to screen all case notes (for example, under circumstances where routine screening is identified by an organisation as a priority or in a search for ADEs). Routine automated screening could also be done within a very short time of an event, conceivably within a short enough timeframe for the patient still to be under active management of an event.

The initial trigger tool concept applied to electronic records, therefore posed a potentially cost effective and timely method for screening large numbers of case notes, of a range of patients and conditions. For example, Jha et al[3] reviewed care over 21964 patient days and compared the results of voluntary reporting of ADEs with chart review (398 ADEs) and computer monitoring using a trigger tool (2620 alerts of which 275 ((99.5%)) were ADEs). Little commentary was made on the low specificity of the computerised alert screening method.

Rozich et al[4] subsequently used Classen’s[1] original ideas to develop and test a more broadly based ADE trigger tool of 24 criteria which could be used across a wide range of hospital types, from community to tertiary hospitals.

Trigger tools are essentially composed of a set of review criteria that are designed to identify possible active incidents or errors, such as the patient being given the wrong medication or a failure by the clinical team to respond to deteriorating vital signs. In this sense, although the evidence base on which they are built may be different, trigger tools are similar to evidence based review criteria derived from clinical practice guidelines. Review criteria that are guideline based are more usually focused on positive acts than are ‘incident based’ trigger tools. For instance, the lack of a guideline based review criterion in the case notes may indicate that an event has not happened – possibly linked to a ‘failure to provide’ event. For example, the failure to record that a measure of glucose level has been taken in a person with diabetes might indicate a more general failure to actively manage the case.

Initially, trigger tools were used either in near ‘real time’ or retrospectively, from paper based or electronic records, to identify possible preventable events in medication safety. Trigger technology was subsequently broadened to identify possible harm in services such as paediatric intensive care, [5] where situation specific trigger criteria have been developed to screen for some of the more frequently occurring, preventable, safety events. In the initial validation study of this tool the most commonly identified adverse events were healthcare associated infections, catheter infiltrates and unplanned extubations requiring re-intubation. Trigger tools for adult intensive care have included such criteria as abrupt falls in haemoglobin level – indicative of severe bleeding or the occurrence of a case of pneumonia in a person who is already a patient in a hospital .[6]

Recent developments have taken a different approach in looking globally for adverse events across whole hospital inpatient systems[7] and providing a measure for comparing one hospital’s results with another, using denominators such as:

Adverse events per 1000 patient days or

Adverse events per 100 admissions or

Percentage of admissions with an adverse event.

More controversially, the Institute of Healthcare Improvement [8] has developed an outpatient care trigger tool that “bands together multiple episodes of care across a continuum” [8], using triggers data from malpractice claims to categorise outpatient care related adverse events. The tool comprises 11 criteria “to provide ‘clues’ to the possibility of adverse events in a patient record” [8].

**Strengths and weaknesses of trigger tool methods**

Although the value of using the electronic screening trigger programmes has been recognised, it is also apparent that many hospital record systems are still paper based, so that more recent versions of trigger tools have been directed towards supporting the screening of case notes by trained reviewers, using what is effectively a criterion based, explicit, approach. Just like review criteria, trigger tools bring structure to a review, being used as a framework for screening case notes or electronic records and identifying pointers to potential adverse events, which are then explored through full, holistic, case note review.

Although these more broadly based service reviews are now being more widely promoted by the Institute for Healthcare Improvement and through projects in the UK NHS [8], rather less is currently being said about the limitations of the method. These limitations can be seen as:

1] development effort required for the criteria for the trigger tool

2] requirement for casenote review where trigger criteria are found

3] validity and interpretation of the results.

Development resource

Extensive effort is required to develop a set of review criteria that have some evidence base, face validity and reproducibility. Although Resar and colleagues do not indicate how much effort was required to develop the four trigger programmes they outline in their 2003 review article,[2] they indicate in another article that many person hours were required to develop the IHI adverse drug event tool.

To create trigger tools for (types of) adverse events, it would be necessary first to create an initial list of possible adverse events for a clinical condition or care setting – this has been the approach in the IHI global adverse events tool.[7] Once an initial list of trigger criteria are developed then validation and reproducibility testing adds a further burden. For instance, development of the IHI medication tool [4] was undertaken in 86 hospitals and was based on a review of 2837 records. This is a highly resource intensive process and it is unlikely that this level of funding will available often, particularly in the UK, so that trigger tools are likely to be limited in number and for the foreseeable future to be essentially of North American origin.

Use of case note review

During a screening review using trigger tools, a positive finding of any one criterion requires that a full case note review must be undertaken. Resar and colleagues [2] point out that ‘the reviewer must review the use of the trigger in the context of the care documented’. For example, in a medication review an event that appears to be an ADE may be an adverse drug reaction [unpredictable and probably not preventable] rather than an adverse event.

One of the key limitations of trigger technology is that any adverse events not identified by a trigger would be missed[9] unless a general screen is carried out, which defeats the efficiency purpose. Moreover, it could be argued that if it takes circa 15 - 20 minutes[4] to manually scan to identify for one in a set of adverse review criteria, this is about the same length of time as an experienced reviewer might take to undertake a structured implicit review. The IHI trigger tool study[4] used full retrospective case note review with 23 (together with one open) review criteria. Instead of being a screening tool, therefore, the trigger criteria could be seen as part of a mixed explicit/implicit case note review methodology.

Validity and interpretation of the results

When trigger tools are developed using rigorous methods and with extensive validation, there is undoubtedly a role for such review methods when they are used to review sets of case notes within an institution and, in combination with full implicit review, they can also be used to explore safety and quality between institutions. However, Brown and colleagues[10] point out a number of methodological limitations when trigger results alone are used as screening methods. Under such circumstances, the arguments comprise concerns over both sensitivity and specificity.

The authors identify the problem of a lack of a ‘gold standard’ for identifying the actual level of events occurring (even observation is not an accurate measure) so that the sensitivity of a specific trigger tool - for example for measuring the rate of ADEs in a particular population – may be higher than reporting but is not as high as some other (possibly more expensive) methods such as full holistic review. Trigger tools can, of course, be used denominator free or used with a denominator such as 100 bed days, just the same as in criterion based clinical audit.

If the specificity of the criteria in a trigger tool is high then only a narrow range of events may be identify. Conversely, Brown and colleagues point out,[10] if specificity is low then there will be many false positives and resource inefficient review. Use of tools with low levels of specificity might yield biased data in comparisons between organisations.

**Is there a role for trigger tools in case note review of care for conditions such as COPD and heart failure?**

The application of trigger tools to care for chronic conditions is certainly possible, in somewhat similar manner to the production of evidence based review criteria. In the evaluation of safety, trigger tools have some advantage in that they can be developed to directly identify possible poor care, unlike the more usual review criteria which usually indicate possible gaps in care. However the extent of the development work required to identify the range of indicators needed to trace possible flaws in care might be even greater than that needed for evidence based review criteria to evaluate the process of care delivery.

It is likely, therefore, that only limited trigger tool sets of criteria will be available in the foreseeable future and that these will be more likely to be applied to specific instances, such as medication events, or to complex and event prone settings such as intensive care units.

Nevertheless, where trigger tool criteria do exist it may be worth exploring their use as electronic medical records become commonplace in hospitals. While methodological limitations will remain, and care will be required in interpreting data from trigger tools that are used to provide ‘comparable’ data, research should be undertaken on the utility of a combined method of trigger tool screening with structured holistic review of identified records.

**References**

1] Classen D, Pestotnik SL, Evans RS, Burke JP. Computerised surveillance of adverse drug events in hospital patients. *JAMA* 1991;266:2847-2851.

2] Resar RK, Rozich JD, Classen D. Methodology and rationale for the measurement of harm with trigger tools. *Qual Saf Health Care* 2003;12: ii39-ii45.

3] Jha AK, Kuperman GJ, Teich JM, Leape LL, Shea B, Rittenberg MA, Burdick E, Seger DL, Vliet MV, Bates DW. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *JAMIA* 1998;5:305-314.

4] Rozich JD, Haraden CR, Resar RK. Adverse drug event trigger tool: a practical methodology for measuring medication related harm. *Qual Saf Health Care* 2003; 12:194-200.

5] Sharek PJ, Horbar JD, , Mason W, Bisarya H, Thurm CW, Suresh G, Fray JE, Edwards WH, Goldmann D, Classen D. Adverse events in the neonatal intensive care unit: development, testing, and findings of an NICU-focused trigger tool to identify harm in North American NICUs. *Pediatrics* 2006;118:1332-1340.

6] Resar RK, Rozich JD, Simmonds T, Haraden CR. A trigger tool to identify adverse events in the intensive care unit. *Joint Commission Journal of Quality Improvement*. 2006;32:585-590.

7] Griffin FA, Resar RK. IHI global trigger tool for measuring adverse events. IHI Innovation Series white paper. Cambridge MA: Institute of Healthcare Improvement; 2007.

8] <http://www.IHI.org> accessed 13 March 2008.

9] Dean B. Adverse drug events: what’s the truth. *Qual Saf Health Care* 2003;12:165-166.

10] Brown C, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, Lilford RJ. *An epistemology of patient safety research: a framework for study design and interpretation*. Birmingham: MRC Patient safety Methodology Network, 2007.

**10.6 Publication 6**

**Hutchinson A**, Coster JE, Cooper K L, Pearson M, McIntosh A, Bath P A. A structured judgement method to enhance mortality case note review: development and evaluation. BMJ Quality and Safety2013; doi:10.1136/bmjqs-2013-001839

**A structured judgement method to enhance mortality case note review: development and evaluation**

Allen Hutchinson, Section of Public Health, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

Joanne E Coster, Section of Public Health, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

Katy L Cooper, Section of Public Health, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

Michael Pearson, Department of Clinical Evaluation, University of Liverpool, Liverpool, UK

Aileen McIntosh, Section of Public Health, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

Peter A Bath, Information School, University of Sheffield, Sheffield, UK

Corresponding Author

Professor Allen Hutchinson

Public Health, ScHARR, University of Sheffield

Regent Court, 30 Regent St. Sheffield, S1 4DA, United Kingdom

Email; [allen.hutchinson@sheffield.ac.uk](mailto:allen.hutchinson@sheffield.ac.uk)

Telephone: 01433650919, Fax: 01142724095

**Keywords**

**Chart review methodologies, Quality improvement methodologies, Healthcare quality improvement, Patient safety, Qualitative research**

**ABSTRACT**

**Background**

Case note review remains a prime means of retrospectively assessing quality of care. This study examines the value of using a new implicit judgement method, combining structured reviewer comments with quality of care scores, to assess the care of people who die in hospital.

**Methods**

Using 1566 case notes from 20 English hospitals, 40 physicians each reviewed 30-40 case notes, writing structured judgement-based comments on the care provided within three phases of care, and for care overall, and scoring quality of care from one [unsatisfactory] to six [very best care].

Quality of care comments on the 119 people who died (8% of the cohort total) were analysed independently by two researchers to investigate how well reviewers could provide structured short judgement notes on quality of care, together with an appropriate care score. Consistency between the explanatory textual data and the related scores was explored, using the overall care score to group cases.

**Results**

Physician reviewers were able to make informative, clinical-judgement based, comments across all phases of care and were usually able to provide a coherent quality of care score relating to each phase. The majority of comments (83%) were explicit judgements. About a fifth of the patients were considered to have suffered from less than satisfactory care, often suffering a series of adverse events.

**Conclusion**

This study shows that a combination of implicit judgement, explicit explanatory comment and related quality of care scores can be used effectively to review the spectrum of care provided for people who die in hospital. The method can be used to quickly evaluate deaths so that lessons can be learned about both poor and high quality care.

**BACKGROUND**

Hospital death rates are a matter of public concern in the United Kingdom and have been the subject of both country-wide data analysis and local intensive reviews, one of which has resulted in a major public debate.[1,2] Concerns about hospital deaths in well-developed health systems, especially when linked to the occurrence of adverse events,[3] have also been expressed internationally. This has resulted in a number of rigorous epidemiological studies of adverse event frequency, for example in Australia, Canada and Sweden.[4-6] More recently there have been large studies of hospital deaths, together with associated events, which have examined whether some hospital deaths might have been preventable. [7,8] On a day to day level, however, there remains a need for rigorous methods to enable clinical teams to retrospectively assess quality of care in a timely manner and, thus, to identify when deaths were inevitable, or whether they might have been prevented with better care. This could assist, for example, in the discussions on care that currently take place in hospital Morbidity and Mortality meetings.

Internationally, case note review remains a prime means of retrospectively assessing quality of care,[3-8] despite the known methodological and practical challenges of this review method.[9-11] Two principal review methods are used:- explicit criterion-based methods, and implicit (sometimes called holistic) methods which are based on clinical judgement.

Criterion-based methods, usually using frameworks of pre-determined criteria to identify elements of care which are either met or not met, are useful for large-scale audits of care or for screening case notes using criterion-based trigger tools.[9]

Implicit review methods are based on clinical judgement, and are probably more effective for identifying and recording the detail and nuance of care (both unsatisfactory and good).[12] Thus, implicit review methods are probably more appropriate for detailed exploration of the care for people who die in hospital. However, unstructured implicit review formats have been criticised for low inter-rater reliability (high variability) and for potential reviewer bias,[9-11,13] whereas structured implicit review limits the variability and creates specific frameworks so that reviewers are able to make, justify and organise statements on care.[14]

Initial models of structured implicit review methods were actually a fusion of implicit judgements of quality of care which were required of the reviewer in order for them to complete a set of explicit review criteria (for example, a criterion such as ‘no appropriate nursing interventions carried out’).[14] A framework such as this was used by Pearson et al to monitor nursing care quality.[15] More recently, Hogan et al used this approach in a study of the frequency of adverse event rates and preventable deaths in English hospitals, where a judgement-based, structured, explicit one-to-five scale was used by reviewers to rate quality of care, from very poor to excellent.[8] In a study of adverse event frequency and preventability upon 8400 patient records in the Netherlands, Zegers and colleagues used two six-point scales on which reviewers used their judgement to record whether injury was caused by health care management or the disease process and to assess the degree of preventability.[7,16]

However, this form of judgement-based structured implicit review only provides a scale-based quantitative result and there is no means of understanding how or why the reviewer judgement was made. Thus the method is useful for large scale monitoring or epidemiological studies of adverse events, but has rather less value for more detailed review at the ward or hospital level of why an event occurred.

To increase the value of structured implicit review in the context of reviewing the whole spectrum of care quality, rather than focussing only on adverse event rates, we designed and tested a structured care review method, drawing on the initial work of Kahn and colleagues.[14] This required reviewers to make implicit clinical judgements and to write explicit comments to support judgement-based quality of care scores.[9] In the developmental stage of the study, multi-professional groups of reviewers independently reviewed the same records, first using a quantitative and then using a qualitative review process. For each case the review process was undertaken for three phases of care (admission, initial management and later management), followed by an overall judgement of the care provided for the patient.For each phase of care, and for care overall, reviewers, both physicians and nurses, were asked to rate quality of care on a one (unsatisfactory) to six (excellent) scale. This was similar to a four-stage, phase of care approach, together with overall care quality, subsequently used by Hogan et al to provide a framework on which to rate quality of care.[8]

There was moderate inter-rater reliability of these judgement-based scores when two or three physicians, working separately, used structured implicit review on the same set of case notes (Intra-Class Correlation Coefficient [ICC] 0.52). Physician reviewers tended to make more explicit written judgements on the quality of care provided than did nurse reviewers, who more often made commentaries about the process/pathway of care.[13]

Subsequently we asked 40 physician reviewers to undertake this enhanced form of structured implicit review to examine the quality of care provided for 1566 people with either Chronic Obstructive Pulmonary Disease [COPD] or heart failure as their main diagnosis. There was no oversampling of deaths and each set of case notes was reviewed only once. There were two reviewers (one for COPD cases and one for heart failure cases) for each of twenty randomly-selected large hospitals in England and each reviewer judged between 30-40 consecutively selected sets of case notes and associated clinical records in their own hospital. Reviewers were either senior respiratory or cardiology physicians in training. Our initial quantitative analysis, reported elsewhere, examined the range of phase of care scores and overall care scores for each of the 20 hospitals and the relationship of the care scores to broader quality of care markers.[9]

Here we report a new qualitative and quantitative analysis of the commentaries written by the reviewers to support their judgement scores of care provided for the 119 cases who died in hospital (8% of the total within the cohort of 1566 cases). The purpose of the analysis was to explore whether physician reviewers can consistently provide short, structured, judgement-based comments on quality of care that they can also justify with an appropriate care score. The consistency between the explanatory textual data and the related scores is explored with a view to considering whether this structured method, combining implicit judgements supported by explanatory comments, together with quality of care scores, can be used for routine mortality case note review.

**METHODS**

**Hospital and reviewer selection**

Acute care hospitals in England were first grouped into quartiles using mortality data. Equal numbers of hospitals from the upper and lower quartiles were then randomly selected (20 in total). Each randomly selected hospital had to provide two reviewers, who were all volunteers and specialists in training. Each was initially approached by specialists in their own hospitals and initial research team contact with the specialists was made through the Royal College of Physicians.

**Reviewer training**

All reviewers received training in the review methods and in data recording prior to data collection. A full-day training session comprised a description of the methods, discussion about the need to be as explicit as possible about the judgement commentaries and a session of reviewing a set of case notes in pairs with tutors. Finally, all of the reviewers judged the care from the same set of anonymised case notes and then commented on their findings in a managed small group discussion, which again emphasised the need to be explicit in their judgements. Data was collected via an electronic form which enabled direct entry by reviewers of both comments and scores for all relevant care phases and care overall. This enabled reviewers to structure their commentaries and the data collection program was also demonstrated during the training day.

Finally, reviewers were provided with a set of national clinical practice guidelines relevant to their clinical specialty. Regular contact was maintained between the study team and the reviewers, who could ask for advice during the review period, using a telephone helpline.

**Data collection**

Each set of case notes was reviewed by a single physician reviewer. Quality of care was assessed in three phases – admission, initial management and later management and also for care overall. For each phase of care and for care overall, reviewers wrote short textual comments on the quality of care provided and were encouraged to be explicit in their comments on care. They also gave the care a score from one to six for each phase and for overall care, based on the criteria in box 1.

**Box 1**

|  |  |
| --- | --- |
| Score 1 | Unsatisfactory: care fell short of current best practice in one or more significant areas resulting in the potential for, or actual, adverse impact on the patient |
| Score 2 | Care fell short of current best practice in more than one significant area, but is not considered to have the potential for adverse impact on the patient |
| Score 3 | Care fell short of current best practice in only one significant area, but is not considered to have the potential for adverse impact on the patient |
| Score 4 | This was satisfactory care, only falling short of current best practice in more than two minor areas |
| Score 5 | This was good care, which only fell short of current best practice in one or two minor areas |
| Score 6 | Very best care: this was excellent care and met current best practice |

**Analysis methods**

Of the 1566 cases reviewed, 119 cases had died during their hospital admission.To explore the type and content of written comments by the reviewers on each of the 119 cases, a textual analysis framework, developed during the study prior to this analysis and previously reported,[9]was applied to all of the phase and overall care comments. Two authors (AH, JC) reviewed and categorised the comments independently and any differences in categorisation were resolved through discussion.

Comments were categorised into three groups (see box 2):

**Box 2**

|  |
| --- |
| Category A. Little or no comment about care and/or little or no judgement, including, for example, a description of what was in the case note or a description of what happened to the patient (not the care they received).  *Note: Category A did not contribute to the analysis presented here, since this analysis was concerned with judgements, rather than descriptive reports.*  Category B. Limited comment about quality of care and/or implied judgement. This category included an implied judgment and/or a description of the care delivered (not just a description of a patient pathway) and/or a description of an omission of care.  Category C. Comments about care with explicit judgements and views. This category included explicit judgements of care delivered, questioning or queries about the care delivered, explanations or justification of care delivered, alternative options or justification of care that should have been delivered or concerns about care. |

All comments in categories B (implicit judgement comments) and C (explicit judgement comments) were subsequently classified by the two study analysts as indicating good quality of care (positive comments) or as indicating poor quality of care (negative comments). These two categories of comment for each case were then grouped by their related overall quality of care scores, which were then used to classify each case into one of six groups, from unsatisfactory care (score one) to very best care (score six). Examples of the detailed textual analysis are presented in the results in Boxes 3-4.

The association between the quality scores for care overall for the group of 119 people who died was compared with the distribution of scores for the 1447 cases who survived, using the Chi-squared test. The association between the comment category and type and their relationship one to another were explored across overall care scores using the Chi-squared test. Chi-squared tests were undertaken using Microsoft Excel and p-values were calculated using Graphpad software (http://graphpad.com/).

**RESULTS**

The overall quality of care scores for the patients who died are compared, in Table 1, with the scores for all patients who survived. The proportions of cases in which care fell short of good practice are relatively similar across the two groups of cases, although there are a higher proportion of ‘satisfactory’ cases and a somewhat lower proportion of the ‘good’ cases among those people who died than in the survivor group. There were no statistically significant differences between the two groups (χ2=9.800; degrees of freedom (df)=5; p=0.0811).

**Table 1 Quality of care overall: score comparisons between people who died and those that survived**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Quality of care** | **Care fell short of good practice** | | | **Satisfactory care** | **Good or better care** | | **Total numbers of reviews** |
| Quality of care scores | 1 (unsatisfactory) | 2 | 3 | 4 | 5 (good) | 6 (very best care) |  |
| People who died (%) | 5  (4.3) | 3  (2.6) | 13  (11.1) | 32  (27.4) | 40  (34.1) | 24  (20.5) | 117\*  (100%) |
| People who survived (%) | 59  (4.1) | 97  (6.7) | 134  (9.3) | 288  (19.9) | 640  (44.2) | 229  (15.8) | 1447\*  (100%) |

(χ2=9.800; degrees of freedom (df)=5; p=0.0811).

\*Two cases from the group of 119 people are not included in this analysis due to incomplete data. Both had phase scores of five or six with no negative comments but for each the overall care score was missing, so they could not be grouped by overall care score.

**Relationship of positive and negative comments to overall care scores**

Table 2 summarises the relationship between the overall care score for each case and the types of comment (whether positive or negative judgements) provided by the reviewers for each of the phases and for overall care. There was a significant association between the total number of positive and negative comments and the overall scores (χ2=205.50; df=5; p<0.0001).

In the care score range unsatisfactory (one) to falling short of best practice (three), the proportion of negative comments outweighs the positive comments. When the care is rated from satisfactory (four) to very best care (six) the positive comments increasingly outweigh the negative. Generally, the positive to negative ratio of comments for each phase remains stable across each overall group score band. So where the overall score is three or less, across each of the phases there are more negative comments than there are positive comments, and the reverse is true for the summary of the higher scores, indicating that the reviewer judgements are generally consistent with the overall score that was given. The ratios of positive to negative comments ranges between 0.28 for overall care score one to 21.17 for those cases grouped by overall care score six.

**Table 2** **Numbers of positive and negative comments per overall score**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overall score  (N cases) | Admission phase comment type and nos. | | Early management phase comment type and nos. | | Later management phase comment type and nos. | | Overall care comment type and nos. | | Total number of comments | | Positive/ negative comment ratio |
|  | Pos | Neg. | Pos | Neg. | Pos | Neg. | Pos | Neg. | Pos | Neg. |  |
| 1 (5) | 6 | 15 | 2 | 18 | 3 | 5 | 2 | 8 | 13 | 48 | 0.28 |
| 2 (3) | 4 | 6 | 2 | 9 | 0 | 5 | 0 | 6 | 6 | 26 | 0.23 |
| 3 (13) | 11 | 19 | 5 | 16 | 6 | 11 | 7 | 14 | 29 | 60 | 0.48 |
| 4 (32) | 39 | 27 | 27 | 24 | 22 | 12 | 24 | 21 | 112 | 84 | 1.33 |
| 5 (40) | 49 | 15 | 49 | 12 | 29 | 2 | 47 | 9 | 174 | 38 | 4.58 |
| 6 (24) | 36 | 4 | 40 | 1 | 25 | 1 | 29 | 0 | 127 | 6 | 21.17 |
| Overall (117) | 145 | 86 | 125 | 80 | 85 | 36 | 109 | 58 | 464 | 262 | 1.78 |

(χ2=205.50; df=5; p<0.0001)

There are fewer comments in total in the later phases of care because some patients died early in the course of the admission. There is also some indication in the textual commentaries that a number of reviewers felt most of what needed to be said had already been said in the earlier phase of care comments for a particular case, and so did not need to be repeated.

In general, the phase of care comments were more detailed than the overall care comments. Occasionally, however, reviewers gave an unexpectedly high score related to a qualitative judgement that suggested a lower quality of care had occurred (see, for example, the case in Box 4).

**Categorisation of comments: implicit and explicit judgements about care quality**

Table 3 summarises the numbers of comments grouped by category (category B – implicit judgements of care, category C – explicit judgements of care) and comment type (positive or negative) for each overall care score.

**Table 3** **Comments by type and category and overall score**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overall score (N cases) | Admission phase comment type, category and number | | | | Early management phase comment type, category and number | | | | Later management phase comment type, category and number | | | | Overall care comment type, category and number | | | |
|  | Pos | | Neg. | | Pos | | Neg. | | Pos | | Neg. | | Pos | | Neg. | |
|  | B | C | B | C | B | C | B | C | B | C | B | C | B | C | B | C |
| 1 (5) | 2 | 4 | 11 | 4 | 0 | 2 | 7 | 11 | 2 | 1 | 1 | 4 | 0 | 2 | 0 | 8 |
| 2 (3) | 3 | 1 | 4 | 2 | 0 | 2 | 4 | 5 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 6 |
| 3 (13) | 2 | 9 | 8 | 11 | 1 | 4 | 7 | 9 | 2 | 4 | 1 | 10 | 1 | 6 | 1 | 13 |
| 4 (32) | 1 | 38 | 2 | 25 | 2 | 25 | 6 | 18 | 0 | 22 | 1 | 11 | 0 | 24 | 1 | 20 |
| 5 (40) | 9 | 40 | 6 | 9 | 2 | 47 | 2 | 10 | 6 | 23 | 0 | 2 | 3 | 44 | 1 | 8 |
| 6 (24) | 9 | 27 | 4 | 0 | 7 | 33 | 0 | 1 | 3 | 22 | 0 | 1 | 3 | 26 | 0 | 0 |
| Total  (117) | 26 | 119 | 35 | 51 | 12 | 113 | 26 | 54 | 13 | 72 | 3 | 33 | 7 | 102 | 3 | 55 |

Results in Table 4 show that, overall, there were more than four times as many explicit comments (judgements) as there were implicit comments. For the lower overall care scores (one to three) there tended to be a rather higher ratio of implicit (B) judgements than there were for the higher care scores, although the implicit judgements were always in the minority. This tend is confirmed by a significant statistical association between the total number of implicit/explicit judgements of care and the overall care score (χ2=48.37; df=5; p<0.0001). Thus the pattern of more explicit comments than implicit comments was seen for all quality of care scores, from one (poor care) to six (best care), indicating that reviewers were on the whole prepared to make explicit judgements where care was poor as well as where care was good.

These results suggest that the reviewers were on the whole prepared to make the type of judgements and explicit comments asked of them during training and which would be valuable in a quality of care review.

**Table 4** **Comparison between implicit/explicit and positive/negative comments**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overall score (N cases) | Ratio of explicit (C) to implicit (B) comments | | | | | | | | |
|  | Total positive comments | | | Total negative comments | | | All comments | | |
|  | B | C | Ratio C to B | B | C | Ratio C to B | B | C | Ratio C to B |
| 1 (5) | 4 | 9 | 2.25 | 19 | 27 | 1.42 | 23 | 36 | 1.57 |
| 2 (3) | 3 | 3 | 1.00 | 8 | 18 | 2.25 | 11 | 21 | 1.91 |
| 3 (13) | 6 | 23 | 3.83 | 17 | 43 | 2.53 | 23 | 66 | 2.87 |
| 4 (32) | 3 | 109 | 36.33 | 10 | 74 | 7.40 | 13 | 183 | 14.08 |
| 5 (40) | 20 | 154 | 7.70 | 9 | 29 | 3.22 | 29 | 183 | 6.31 |
| 6 (24) | 22 | 108 | 4.91 | 4 | 2 | 0.50 | 26 | 110 | 4.23 |
| Total (117) | 58 | 406 | 7.00 | 67 | 193 | 2.88 | 125 | 599 | 4.79 |

χ2=48.37; df=5; p<0.0001

**Content and nature of comments**

Study of the individual comments showed that a number of B category comments contained concise technical summaries in addition to implicit judgements on the quality of care. Many of the C category comments were incisive clinical observations with a strong view of the quality of care, especially when the reviewer considered that the care was poor. Comments across the range of overall scores often included consideration of the broader, non-technical, processes of care (for example, communication with relatives), as well as technical aspects of care.

Of the 21 case reviews with low overall scores (scores of one, two or three), 15 were accompanied by an explicit clinically-relevant judgement that justified the low score. Some related to cases where care was generally poor throughout the inpatient episode, while others related to cases where a specific aspect of care was of concern. In two of the cases, incorrect diagnosis was the main problem, while in 12 cases there was concern about suboptimal management. There were usually multiple smaller events that were additive, rather than one main adverse event, which only occurred in one of the 12 cases. Two of the 15 cases were considered to have such poor record keeping as to be a threat to the care of the patient.

Boxes 3-4 demonstrate some of the range, type and category of comments made by reviewers in two cases. All of the comments are as written by the reviewers and the scores given for each phase of care are included. Reviewers were able both to comment on the technical aspects of care and to take a holistic view of the overall care plan.

Box 3 is also used to demonstrate how the categorisation of the comments was applied in the analysis. For example:

* Although the reviewer explicitly grades the documentation as poor in the admission, this is only a description of the documentation without any explanation and therefore is categorised as a B level comment. In the initial management phase, however, there is a judgement (very poor documentation) together with an explanation, which rates a C category.
* When the reviewer implies in the initial management phase that it was poor practice not to take an arterial blood gas sample (‘No ABGs (Arterial Blood Gas) and patient was tachypnoeic and hypoxic’) there is no explicit statement that this was unsatisfactory (and it is thus a B category comment).
* A judgement on the therapy (‘pitiful dose of frusemide (furosemide) (20mg Intravenous)) is a C category comment.
* When commenting on the technical aspects of care, the reviewer could also be explicit about how the care should have been managed overall, in the context of the patient’s illness. This is an explicit, category C, judgement.

The case in Box 3 also illustrates a pattern that is of a group or ‘constellation’ of events which of themselves may not cause severe harm but which, taken together, can lead to harm to the patient. This pattern was also found in the main study among some of the patients who survived.[17]

**Box 3**

|  |  |  |
| --- | --- | --- |
| Overall care score 1 |  | Comment type (Pos or Neg) and category |
| Admission  Phase score 1 | Poor history documentation  Poor examination documentation  Initial investigations requested CXR (Chest radiograph), ECG (Electrocardiograph), bloods but no comment made re these  No ABGs (Arterial Blood Gasses) and patient was tachypnoeic and hypoxic.  No O2 (Oxygen) [not documented]  No GTN (Glyceryl trinitrate,  pitiful dose of frusemide (furosemide) (20mg IV (Intravenous))  Extremely poor management | Neg/B  Neg/B  Neg/B  Neg/B  Neg/B  Neg/B  Neg/C  Neg/C |
| Early management  Phase score 1 | Medical team made no attempt to adequately treat the heart failure  No comment on the CXR  No ABGs  CPAP (Continuous positive airway pressure) started without ABGs  Did record a resuscitation status  Documentation very poor e.g. no reference to the fact that she was so unwell or whether they thought it likely that she would die.  No discussion with the family or relatives | Neg/C  Neg/B  Neg/B  Neg/B  Pos/B  Neg/C  Neg/C |
| Later management | See previous |  |
| Overall care Score 1 | All aspects of this case were very poor. History, examination, medical management, documentation.  If this lady was clearly dying and had multiple co-morbidities, they should have documented this, made the lady comfortable and called the family in. | Neg/C  Neg/C |

Although there are usually more negative comments than there are positive comments when overall care scores are low, as shown in table 2, the case in Box 4 shows examples of how positive and negative comments can be juxtaposed in each phase. In retrospect, this case also brings the question of whether the overall score of three was the most appropriate – it might be argued from the level of the comment that the case could have been given a lower overall care score of two (see, for example, the comments on later management).

**Box 4**

|  |  |  |
| --- | --- | --- |
| Overall care score 3 |  | Comment type (Pos or Neg) and category |
| Admission  Phase score 4 | ph 7.436  Good history taken of COPD (Chronic Obstructive Airway Disease) symptoms and normal functional status,  alternative diagnosis of PE (Pulmonary Embolus) & CCF (Congestive Cardiac Failure) not excluded in a patient with risk factors for both.  Clinical cardiovascular exam examination not thorough (no mention of JVP (Jugular Venous Pressure), pedal oedema, chest expansion, sputum characteristics) | Pos/C  Neg/B  Neg/C |
| Early management  Phase score 4 | Patient received appropriate treatment for COPD (i.e. steroids, antibiotics and nebulizers),  however the CXR result was never recorded ?looked at. | Pos/C  Neg/B |
| Later management Phase score 2 | Although the patient was recorded to be clinically improving 2/7 post admission and team were considering early discharge, his ABG was not improving and patient's SOB (Shortness of breath) + tachypnoea attributed to anxiety, pt (patient) gradually deteriorated.  Patient changed to inhalers too soon.  Seen appropriately by respiratory team, frusemide (furosemide) and aminophylline infusion appropriately suggested.  Nursing staff inappropriately withheld oral medications as they thought he was nil by mouth.  Developed severe type 2 respiratory failure but no decision on resus (Resuscitation) status made until patient very unwell. This needed to be made by on call team.  Earlier referral to ITU (Intensive Therapy Unit) and I.v aminophylline may have changed outcome.  Good chest physio(therapy) input. | Neg/B  Neg/C  Pos/C  Neg/C  Neg/B  Neg/C  Pos/C |
| Overall care  Score 3 | Patient appropriately treated initially with nebs (nebuliser), antibiotics and steroids  however patient's treatment plan not escalated until he was in severe type 2 respiratory failure.  NIV (Non invasive intubation)/ITU not considered in this patient ?why-he had no other co-morbidities and no previous hospital admissions.  Resus decision made inappropriately by on call team when patient very unwell. | Pos/C  Neg/B  Neg/C  Neg/C |

Comments on good care tended to be more global than those for unsatisfactory care but may also be quite explicit. Cases which demonstrate this and also how a single adverse event may change the reviewer’s overall consideration of the case are included as additional material. (*see Box 5 and Box 6* *attached at the end of the paper for BMJ.com*)

Some of the reviewers in this study were more ‘explanatory’ than others so that, in some cases, the number of comments may reflect individual style rather than the strength of the comment. For example, comments such as ‘good care’ or ‘unclear treatment’ are short explicit judgements without further detail, while other reviewers are more extensively explicit.

Of the 63 case reviews (54% of the total number of mortality reviews) that scored most highly (5 or 6), there were 52 accompanied by a short explicit comment in the overall care section indicating that all key aspects of care had been good or excellent (e.g. ‘well looked after’) and in 16 of the 63 reviews there were comments about the inevitable outcome of the case despite the good care received.

**DISCUSSION**

In this study we have shown that physician reviewers are able to use structured review to, make implicit quality and safety judgements, write explicit short care commentaries and give coherent matching quality of care scores. Quantitative scores and qualitative comments matched well, indicating that physician reviewers can appropriately score the quality of care on a rating scale.

These physician reviewers could identify and explain both technical and non-technical aspects of care, and could rank these aspects of care using a set of ‘benchmark’ scores, ranging from very good care to very unsatisfactory care. For people with complex illnesses, the outcome is not always survival. However, structured explicit judgments can show how high quality care was provided – even if the patient has not survived. For example, there were a number of instances where explicit comments were made about the quality of non-technical care such as the way information was provided to patients and their relatives. Conversely, when poor care occurs, the method can identify the points at which care fails to meet expected standards, and when the situation can be, or is, rescued. It is interesting to note that in Table 1 the proportions of those who died and had less than satisfactory care (about 20% of the cases), were similar to those who survived and had poor care.

During the training session, reviewers were encouraged to be as direct as possible in their commentaries, and in the results overall (Tables 3 and 4) there were many more explicit comments than there were implicit comments. Nevertheless when poor care was being described, while explicit comments predominated, there was a noteworthy proportion of implicit, B level, comments. Sometimes these B level comments were about documentation (which was not in the C category) or concerned missed tests which the reviewer listed and didn’t specifically make a judgement upon – for example, No ABGs (Arterial Blood Gases), see Box 3. It may be that in this case the reviewer felt that the result said it all and that an explicit comment was superfluous. On the other hand, it could also be that some reviewers might have felt uncomfortable about making direct comments about very poor care.

With the hindsight of these results, and when undertaking reviews such as this in health service settings, training should include discussion of an initial sample of commentaries and scores with each reviewer to assist in maximising the number of explicit comments. Of course, training might identify some reviewers who do not feel able to make explicit comments and so would not be suitable for this type of review.

The phase of care structure also contributes to an understanding of how care may vary, and at what point. Interestingly, a phase of care approach has also been used by Shannon and colleagues in a review of cardiac surgical care,[18] albeit in a rather more structured system with distinct changes in physical settings. In the context of assessing whether death was a preventable outcome, Hogan et al used a four-phase model to identify adverse incidents – initial assessment, treatment plan, ongoing monitoring and preparation for discharge.[8] Under the conditions of a service review a three-phase model might be easier to manage but either a three or four phase approach would be appropriate.

Qualitative comments from the reviewers were useful in that they could succinctly identify what was done badly in poor cases. Such short explicit judgements could support a wider, more detailed, service review to assess what could be improved in a particular setting or condition. Furthermore, since this structured review method assesses both process and outcome of care, this mixed type of review, using qualitative comments with scores, might be a useful addition to review measures which only assess outcomes or are criterion based. This mixed qualitative and criterion based method is published in detail elsewhere.[9]

In this study, assessments of the quality and safety of the care provided showed that, for over 80% of the patients who died, care was rated at least satisfactory and, for approximately half of the cases, care was judged to be of high quality. The processes of care described enable a qualitative judgement to be associated with an objective score that is explicable to, and understandable by, a wide range of people and would be understood by the public too. However, having graded a case as poor or not, there is the added advantage that the structured comments also provide the reasoning behind the judgement in a format to which clinical teams and individuals should be able to respond in a review process.

**Limitations**

In this study the 40 reviewers were all volunteers who undertook the work in their own hospitals. Although there might be concerns about the impartiality of using internal review teams, results have shown that reviewers can make incisive short notes (commentaries) about quality of care, and can critically review care provided in their own hospitals.

Internal review teams have also been used in other settings. Sharek et alcommented on the strong performance of hospital-based internal review teams, albeit when using more structured, criterion-based, trigger tools to identify adverse events.[19]

Although it could be argued that two reviewers per case might enhance the quality and depth of a case note review, there is some evidence to suggest that this use of a more intensive resource does not necessarily improve the review process. While we were able to show in our development study that there was reasonable coherence of quantitative care scores and criterion-based scores between physician reviewers,[9,13] other work by Hofer and colleaguesfound that multiple reviewing of the same set of case notes did not enhance the results.[20]

Finally, it is important to recognise that there are limits to the extent to which the quantitative analysis of the reviews can be used. For example, averaging phase scores across each case, to determine whether phase score averages are similar to the overall care score, is not appropriate. An example of this can be found in the online resource in Box 6 where care was judged excellent until moments before the patient died. The value of this current study is that the context and the basis for any quantitative score can be found in the phase of care comments associated with each score.

**Conclusions**

This method is a refinement on both global implicit judgement and structured implicit judgement used upon a set of case notes, because it is able to provide information on aspects of each phase of care, enabling more detailed, yet still brief, comments to show explicitly how care may vary or be consistent with expected standards. For example, this method could be used to identify whether care has led to a preventable death, or to identify good quality of care even though the overall outcome is failure to survive. Thus, although the study did not explicitly seek to judge a death as preventable, as did Hogan and colleagues,[8] review training could straightforwardly include an explicit judgement commentary about whether a death was preventable or was not preventable (as some of the study reviewers actually did).

Results also show how explicit written judgements and quality of care scoring can be used together and thus may offer a range of case note review methods for use under differing circumstances, together with opportunities for providing training and assessment of ‘reviewer quality’.

Structured judgement review provides the framework for a quality of care review that can be used by clinical leaders and quality managers to identify potential priority areas for evaluation. For example, scoring allows for a screening of the overall care quality for a case overall, or can identify issues in a particular phase of care, say at admission or initial management. Explicit comments allow exploration of particular aspects of care, for instance where good treatment plans might be inadequately implemented. For these purposes it is not necessary to analyse whether comments are implicit or explicit. The data collection framework is straightforward, has been previously published and is easily available.[9]

Who should act as the reviewers? Because of the complexity of illness often presented in hospital settings, studies of adverse events have used experienced generalists with some specialist support.[8] This structured implicit review method could be used in a similar way either with in-hospital teams or by visiting teams from other hospitals. We do not know whether the review results would be better when undertaken by experienced specialists than the reviewers in our study. However, our results have shown that this form of review can be undertaken by specialists at a senior level in a training programme – so increasing the pool of trained senior reviewers in a hospital - and thus the method offers the opportunity for early review of the care of people who die in hospital so that, where necessary, timely quality improvement lessons can be learnt.

**ABBREVIATIONS**

COPD: Chronic Obstructive Pulmonary Disease

**RESEARCH ETHICS REVIEW**

A research ethics review of the study was sought from the Trent Multi-centre Research Ethics Committee on 21 July 2004, prior to the start of data collection. Because, in both phases, data were to be collected by staff working in each hospital, and the data were anonymised before transmission to the research team, the Committee considered this to be equivalent to a national audit programme. The Trent Multi-centre Research Ethics Committee response was therefore that the study did not require an ethics opinion from the Committee.

**COMPETING INTERESTS**

The authors declare that they have no competing interests.

**AUTHOR CONTRIBUTIONS**

Allen Hutchinson took the lead in the conception and design of the study, took the lead on the analysis of the qualitative mortality review data, was principal author of all of the drafts of this paper and has given approval for this version of the paper to be published.

Joanne Coster contributed to the conception and design of the study, undertook data collection and analysis of the mortality review data, contributed to all of the drafts of this paper and has given approval for this version of the paper to be published.

Katy Cooper made contributions to the design of the study, undertook data collection and contributed to analysis of the mortality review data, contributed to all of the drafts of this paper and has given approval for this version of the paper to be published.

Michael Pearson contributed to the conception and design of the study overall, contributed to the interpretation of the mortality review data, contributed to all of the drafts of this paper and has given approval for this version of the paper to be published.

Aileen McIntosh contributed to the conception and design of the study overall, took a lead on the qualitative analysis framework, contributed to all of the drafts of this paper and has given approval for this version of the paper to be published.

Peter Bath contributed to the conception and design of study and of the qualitative analysis framework, undertook the statistical analysis for the quantitative analysis, contributed to all of the drafts of this paper and has given approval for this version of the paper to be published.

No other authors contributed to this analysis and authorship.

Allen Hutchinson acts as Guarantor for this article.

**ACKNOWLEGEMENTS**

This project was funded by the National Institute for Health Research Health Technology Assessment (NIHR HTA) Programme (project number RM03/JH08/AH) and was published in full in Health Technology Assessment 2010,**14(10):**1-170.

The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the HTA programme, NIHR, NHS or the Department of Health.

**REFERENCES**

1. The Dr Foster Hospital Guide 2009: How Safe Is Your Hospital? London: Dr Foster Ltd 2009.
2. The Stationary Office: Report of the Mid Staffordshire NHS Foundation Trust Public Inquiry. London: The Stationary Office 2013. ISBN 9780102981476.
3. Thomas EJ, Brennan TA. Incidence and types of preventable adverse events in elderly patients: population based review of medical records. BMJ 2000;**320**:741-4.
4. Wilson RM, Runciman WB, Gibberd RW et al. The Quality in Australian Health Care Study. Med J Aust 1995;**163**:458-71.
5. Baker GR, Norton PG, Flintoft V et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ 2004;**170**:1678-86.
6. Soop M, Fryksmark U, Koster M et al. The incidence of adverse events in Swedish hospitals: a retrospective medical record review study. Int J Qual Health Care 2009;**21**:285-91.
7. Zegers M, de Bruijne MC, Wagner C et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care 2009;**18**:297-302.
8. Hogan H, Healey F, Neale G et al. Preventable deaths in care in English acute hospitals: a retrospective case record study. BMJ Qual Saf 2012. doi:10.1136/bmjqs-2012-001159.
9. Hutchinson A, Coster JE, Cooper KL et al. Comparison of case note review methods for evaluating quality and safety in health care. Health Technol Assess 2010;**14**:1-170.
10. Lilford R, Edwards A, Girling A, et al. Inter-rater reliability of case-note audit: a systematic review. J Health Serv Res Policy 2007;**12**:173-80.
11. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA 2001;**286**:415-20.
12. Mohammed MA, Mant J, Bentham L et al. Process and mortality of stroke patients with and without do not resuscitate order in the West Midlands, UK. Int J Qual Health Care 2006;**18:**102-106.
13. Hutchinson A, Coster JE, Cooper KL et al. Assessing quality of care from hospital case notes: comparison of reliability of two methods. Qual Saf Health Care 2010 Dec, **19(6)**:e2.
14. Kahn KL, Rubenstein LV, Sherwood MJ et al. Structured Implicit Review for Physician Implicit Measurement of Quality of Care: Development of the Form and Guidelines for Its Use. California: Rand Corporation 1989.
15. Pearson ML, Lee JL, Chang BL et al. Structured implicit review: a new method for monitoring nursing care quality. Med Care 2000;**38**:1074-1091.
16. Zegers M, de Bruijne MC, Wagner C, et al. Design of a retrospective patient record on the occurrence of adverse events among patients in Dutch hospitals. BMC Health Serv Res 2007;7:27 doi:10.1186/1472-6963-7-27
17. Hutchinson A, McIntosh A, Coster JE et al. When is an event an event? Data from the quality and safety continuum. In: From safe design to safe practice. Edited by Hignett S, Norris B, Catchpole K et al. Cambridge: The Ergonomics Society 2008. ISBN 978-0-9554225-2-2
18. Shannon FL, Fazzalari FL, Theurer PF et al. A method to evaluate cardiac surgery mortality: phase of care mortality analysis. Ann Thoracic Surg 2012;**93**:36-43.
19. Sharek PJ, Parry G, Goldmann D et al. Performance characteristics of a methodology to quantify adverse events over time in hospitalised patients. Health Serv Res 2011;**46**:654-678.
20. Hofer TP, Bernstein SJ, DeMonner S et al. Discussion between Reviewers Does Not Improve Reliability of Peer Review of Hospital Quality. Med Care 2000;**38**:152-61.

**ADDITIONAL MATERIAL**

Box 5 and box 6 provide additional examples of reviewer scores and comments, together with the assessment of each comment type and category. Box 5 is a case where the care was judged to be good. Overall there were fewer comments than those cases where care was judged to be unsatisfactory. In the later management section of this case there is a description of care included which, though not a judgement and therefore not placed in a category, illuminates the case and its management for the reader.

**Box 5**

|  |  |  |
| --- | --- | --- |
| Overall care score 5 |  | Comment type and category |
| Admission  Phase score 5 | Good initial assessment.  Heart failure identified and confirmed radiologically and appropriate treatment commenced.  Important differential diagnoses considered and also treated for chest infection. | Pos/C  Pos/C  Pos/C |
| Early management  Phase score missing | Regularly reviewed as he failed to improve and significant effort made to optimise his medication. | Pos/B  Pos/C |
| Later management  Phase score 5 | Ceiling of treatment considered, documented and communicated to family. When it became apparent that his situation was hopeless, active treatment was withdrawn.  Again, communication with the family was good and documentation was clear. | Pos/C  Pos/C |
| Overall care  Score 5 | Well looked after.  A decent attempt to treat her from a difficult starting point.  Good multidisciplinary involvement and  good communication with the patient and family throughout. | Pos/C  Pos/C  Pos/C  Pos/C |

Box 6 is an example of how the impact of one adverse event may impact on the overall score, even when care is good in other components of care. Just at the point when the patient died, an inappropriate action occurred which the reviewer regarded as very unsatisfactory, although it did not contribute to the patient’s death. The reviewer gave high care scores for three phases but then deemed the overall care poor because of the adverse event at point of death.

**Box 6**

|  |  |  |
| --- | --- | --- |
| Overall care score 1 |  | Comment type (Pos or Neg) and category |
| Admission  Phase score 5 | excellent history taking and clinical examination. Oedema of feet not recorded though is noted in subsequent ward examinations. | Pos/C  Neg/ B |
| Early management  Phase score 6 | appropriate medical management.  Decision taken to manage conservatively as end stage heart failure | Pos/C |
| Later management  Phase score 6 | discussion with family about end stage heart failure. Decommissioned ICD as was appropriate | Pos/C |
| Overall care  Score 1 | Excellent management as deemed appropriate in this terminal situation. DNR form done by team.  Despite this when patient was in periarrest situation, patient was given 1 DC shock. This did not seem appropriate at all as DNR decision was discussed with the family. | Pos/ C  Neg/ C |