Large-scale mixed-methods evaluation of safety programmes and interventions

Graham Martin, Jane O'Hara and Justin Waring

Accepted version of a chapter published in K. Pettersen-Gould and C. Macrae (eds) Inside hazardous technological systems: methodological foundations, challenges and future directions. London: CRC Press, 2021.

6267 words

Abstract

This chapter provides an overview of approaches to evaluation of safety and quality improvement initiatives in the field of healthcare. It traces the history of research and evaluation in this field, noting how a tradition of experimental study design and methodological innovation had endowed it with some enviable features, such as a strong evaluation toolkit, a robust evidence base, and exacting standards that new interventions must meet before widespread adoption. An increasingly sophisticated array of methods has been applied to the evaluation of quality and safety innovations, including mixed-method designs that combine quantitative and qualitative approaches to offer a nuanced understanding of whether and how interventions work. The authors present examples of several mixed-methods evaluations. They offer critical reflections on the state of the field, highlighting in particular how for all their advantages, the high methodological and evidentiary standards that prevail in healthcare also produce down sides for the study of patient safety, where the complexities of both the problems and the vaunted solutions mean that a definitive evidence base—of the kind valorised in health services research—will likely always be elusive.

Introduction

In this chapter we explore the terrain of research and evaluation within safety in healthcare. In contrast to many of the fields explored in this volume, healthcare is distinguished by a long history of methodologically robust research. Indeed research in health and healthcare has often been at the vanguard of methodological development to increase the validity and reliability of results. Both pharmaceutical and non-pharmaceutical interventions to relieve disease or improve health are typically (though not always) subject to an extensive array of evaluation processes before being incorporated into routine healthcare delivery, including for their safety, and they are often relatively well monitored for their continuing safety and effectiveness after adoption. However, this tradition of evaluation, we argue, brings challenges as well as advantages to the field of research and evaluation around quality and safety in healthcare. In particular, the dominance of methods associated with epidemiology and pharmaceutical development has meant that it is only relatively recently that key institutions in the field of health services research have come to recognise the contribution that qualitative inquiry, including ethnography, can make. More lately, however, the field has begun to enthusiastically adopt mixed-methods research, acknowledging the complexity both of safety interventions and the contexts in which they are expected to work, and of the role of qualitative insights in understanding how interventions lead to better safety—and the extent to which a seemingly effective intervention can be expected to work consistently through time and space.

Understanding of the prevalence and nature of issues of patient safety and quality in healthcare is a relatively recent phenomenon. It is only since the 1980s and 1990s that quality

and safety in healthcare has become a field of study and intervention in its own right. Attention to the issue was galvanised by reports on medical error, patient harm and associated poor outcomes (Department of Health 2000; Institute of Medicine 1999), and widely publicised (though disputed) figures that suggested that medical error was among the leading causes of death (see Shojania 2012). Similarly startling and contentious claims have been made regularly ever since. For example, one recent study ranks medical error as the third-highest cause of death in the United States (US), behind cancer and heart disease (Makary and Daniel 2016; see Shojania and Dixon-Woods 2017 for critique of the validity of this claim). The World Health Organization (WHO) (2018) suggests that in high-income countries, one patient in 10 is harmed as a result of adverse events while receiving hospital care. Over the last two decades, research in several countries, as well as analysis of the wealth of data routinely collected in various national audits, registries and billing systems, had identified areas of high risk, poor reliability and inconsistent outcomes, notably in surgery, anaesthesia, and post-operative care. This in turn has given rise to interventions that seek to address these issues, often based on work in other high-risk and safety-critical industries such as civil aviation, and accompanied by often robust, multifaceted efforts at evaluation. While this work has found few if any 'magic bullets', it has contributed to an increasing understanding of the complexities of problems of quality and safety in healthcare—and the need for sophisticated interventions to address them. We discuss some key mixed-methods studies in the field of patient safety, highlight the contributions to understanding they have offered, and examine some of the challenges that continue to face mixed-methods research and evaluation. In particular, we consider some of the issues that have made realising the promise of mixed-methods research difficult, and the prospects for overcoming these issues.

Our chapter is presented in three main sections. First, we briefly recount the history of research in healthcare, tracing its consequences for dominant assumptions about appropriate methodology and epistemology in the study of interventions to improve safety—and the more recent acknowledgement of the role of mixed methods, particularly in evaluating complex interventions. Next, we trace the rise of concerns about safety, quality and risk in healthcare from the 1980s onward, and discuss examples of increasingly sophisticated mixed-methods evaluation of safety interventions in healthcare that emerged in response. We compare a number of programmes and their evaluations, and note both the insights that have been brought by the incorporation of qualitative methods, and some of the limitations that have emerged. We pick up this theme in the final section, where we argue that although qualitative research has undoubtedly brought analytical advantage, some of its potential may have been overstated—due to the challenges of evaluation in a complex field, epistemological differences between researchers, and conflicting expectations of the endeavour of evaluation itself. We conclude by discussing prospects for evaluation in this field (and in other areas of safety and risk), highlighting the need for more modest ambitions for evaluation of safety programmes and greater attention to the relationship between researchers and practitioners.

Research and evaluation in healthcare

To arguably a greater extent than many other safety-critical industries, the field of healthcare has been dominated for some time by the ideals of evidence-based practice—that is, the notion that routine activities should be guided by strong evidence that interventions are likely to achieve what they intend to (Sackett *et al.* 1996). Examples of the use of experimental methods in relation to medical treatments from as early as the sixteenth century have been noted (Oakley

2000), and occasional appeals to the need for evidence can be found from the early twentieth century onward, but the origins of healthcare's contemporary evidence-based practice 'movement' were in the 1970s and 1980s (e.g. Cochrane 1972). Stemming both from scandals in the 1960s that arose from insufficient trialling of pharmaceutical interventions, and from concerns about ensuring the most effective use of scarce healthcare resources (Howick 2011), this movement called for healthcare practice to be informed by high-quality evidence of the effectiveness of interventions deployed. In particular, advocates of evidence-based practice argued that healthcare interventions should be evaluated through fair and objective processes, free of both the prior preferences of researchers and clinicians and the biases of vested interests such as the pharmaceutical industry, and effectively curated by the clinical-academic community to ensure that the best and most up-to-date evidence was readily available to practitioners. These premises can be seen reflected in the so-called 'hierarchy of evidence' (see Figure 1), a model for assessing the validity and reliability of research findings that remains influential in healthcare. This typically places systematic reviews—i.e. integrations of high-quality research evidence, filtered according to explicit inclusion criteria and often incorporating meta-analyses to estimate effect size—at the apex, followed closely by randomised controlled trials—i.e. tests of one intervention against another (or against no intervention), with steps (e.g. randomisation, blinding) taken to eliminate biases arising from differences in sample characteristics, researcher or clinician preferences, and so on.

[Insert Figure 1 about here]

As a set of aspirations, there is little to object to in the evidence-based practice vision, and as a set of tools for producing a high-quality evidence base to inform practice, it is easy to see its appeal. As an alternative to the kind of judgement-based practice that had predominated in medicine for centuries—and which resulted sometimes in cure, sometimes in harm, often in inefficiency and always in inconsistency—it presents an attractive model of medicine as a science, rather than an art reliant on the skill and attention of the individual practitioner. But the limitations of this vision are also apparent. Originators and critical friends of the evidence-based practice movement alike have highlighted how, for example, an unreflexive application of universal rules indicated by an epidemiologically derived evidence base to individual patients who are all unique can itself produce harm, and fail to incorporate wider considerations—not least patients' own preferences (Greenhalgh *et al.* 2014; Sackett *et al.* 1996). But perhaps particularly pertinent to the focus of this volume is the argument that the notion of a universal evidence base, which identifies and calls for implementation of the best single interventions for any given issue, inclines researchers towards a particular set of problems and a particular set of solutions. While placebo-controlled, double-blinded randomised controlled trials may be the most robust way to evaluate a relatively simple pharmaceutical intervention, it is quite another thing to argue that a similar experimental model is the only (or even the best) route to a high-quality evidence base on more complex interventions—i.e. those comprising multiple components, perhaps interacting in unpredictable ways, and with unknown or multifaceted pathways between cause and effect. Further, in many randomised controlled trials, internal validity has traditionally been prioritised over generalisability, with strict participant eligibility criteria to ensure a homogeneous, but often rather exclusive, sample. This has meant that some populations (e.g. white, male) have been better served by the evidence-based practice paradigm than others. Indeed, it could be argued that it has itself introduced risks to the safety and quality of healthcare, since interventions that

work well in clearly defined populations with a single disease may be much less effective—or even, due for example to drug interactions, harmful—in patients with multiple morbidities who increasingly constitute much of the real-world population (Boyd and Kent 2014).

Yet for a long time, study designs elevated by the hierarchy of evidence dominated healthcare research, and it was only towards the end of the twentieth century that alternative epistemological frameworks began to get a hearing in biomedical circles. The 1990s saw engagement of the *BMJ* with qualitative methods (e.g. Pope and Mays 1993). The reorganisation of healthcare research in the United Kingdom (UK) in the 2000s, including the inauguration of the National Institute for Health Research (Department of Health 2006), provided new opportunities for mixed-methods research and evaluation, by providing a platform for large-scale applied health research. At the same time, reflecting realisation of the limitations of the randomised trial model for non-pharmaceutical interventions, the UK Medical Research Council (MRC) introduced a framework for complex interventions which explicitly acknowledged the need for qualitative as well as quantitative methods in their development and evaluation (Campbell *et al.* 2000). Subsequently this framework was revised to better acknowledge the role of complexity and the need to tailor evaluation to specific interventions (Craig *et al.* 2008). Agencies elsewhere, for example in the US (PCORI 2018), have adopted similar sets of standards, and mixed methods are now *de rigueur* in evaluation in healthcare, particularly in relation to multi-component, complex interventions where causal mechanisms, impacts and unintended consequences may vary with context. Alongside this, theory-based evaluation in the mould of Carol Weiss (1995), most prominently realist evaluation (Marchal *et al.* 2012; Pawson and Tilley 1997), has also become common, and seeks to give explicit attention to understanding causal mechanisms as well as outcomes, with a view to understanding how both might vary through time and space.

Mixed-methods evaluation is thus now widely accepted in health services research in the UK and elsewhere, and even expected by its funders. But as we later explore in more detail, realising the promise of mixed methods—in relation to patient safety and other fields of healthcare research—has not always proven easy, in part because of the legacy of the history outlined above and the persistence of epistemological hierarchies, and in part because of the challenges of integrating rather different modes of understanding in a truly synergistic way. First, however, we turn our attention to the rise of patient safety as a concern in healthcare, and the response to this concern in the form of research, intervention, and evaluation.

Patient safety, improvement interventions, and mixed-method evaluation

The emergence of patient safety as a concern for clinicians and health services researchers is in some ways bound up with the rise of evidence-based practice précised above (Waring *et al.* 2016). The development of an increasingly robust (if not always entirely generalisable) evidence base for healthcare interventions was not immediately matched by changes in practice among 'rank-and-file' clinicians. The 1990s and 2000s saw the publication of a number of studies that highlighted the extent to which practice lagged behind current evidence (e.g. McGlynn *et al.* 2003; Schuster *et al.* 2005) and the frequency of medical errors (e.g. Brennan *et al.* 1991), and the consequences of these issues in terms of suboptimal outcomes, wasted resources and patient harm. Policy reports from a similar period highlighted the consequences of variation in healthcare practice and deviation from evidence-based standards, and particularly the issue of patient safety problems arising from increasingly complex healthcare systems, suboptimal design and human error, in the UK and the US (Department of Health

2000; Institute of Medicine 1999). These reports contained notable parallels. Both posited that learning from past problems could have prevented later issues; both drew on insights from human factors and related fields in acknowledging the limitations of human cognition and the need for systems-based approaches to improving safety; and both highlighted the success of other industries in reducing, managing or eliminating safety risks. The combination of concerns around inconsistent quality of care and failures of patient safety gave impetus to a range of activities in healthcare policy, practice and research (see Shojania 2012 for a brief overview). In the UK, this included regulatory interventions such as clinical governance and inspection and audit regimes, the growth of quality-improvement work driven by local audits of clinical practice, and the development of large-scale, cross-cutting programmes to address quality and safety issues, accompanied by similarly ambitious programmes of research and evaluation.

The rapid rise of patient safety concerns from relative neglect to policy priority has, however, meant that improvement practice has often exceeded the research base—in conflict, of course, with the tenets of evidence-based practice discussed above. Various authors have noted that interventions to reduce risk and improve safety have emerged in rather a patchwork fashion, with the field characterised in its early stages by multiple projects springing up, uninformed by existing evidence, with weak bases in theory, and without high-quality evaluation. "Safety initiatives have been promoted […] before robust research had demonstrated the effectiveness of these practices" (Wachter 2010: 169; cf. Dixon-Woods and Martin 2016). More sophisticated, co-ordinated and ambitious attempts to improve safety, recognising the multifaceted nature of safety problems and the need for well-theorised interventions in response, have begun to follow. Thus efforts to improve safety based on a single intervention with a simple theory of change—for example, auditing current practice and feeding back findings to clinicians, premised on the implicit or explicit assumption that poor practice is due to deficits of knowledge or awareness—have been supplemented by much more nuanced and programmatic interventions. These interventions have been more theory-based (drawing, for example, on psychological insights into how to prompt sustained behaviour change), and have sought to account for influences on safety at multiple levels, for example professional norms, patient preferences and organisational incentives (Shekelle *et al.* 2011).

In other words, safety interventions in healthcare have quickly moved from simple to complex. In this way, they have followed the calls of organisations like the MRC discussed in the previous section (e.g. Craig *et al.* 2008), and started to draw on a more sophisticated understanding of the principles of safety science more broadly, with its emphasis on systems issues, which present complex challenges and demand complex solutions (Braithwaite 2018). Research and evaluation, however, have arguably lagged behind. Whether because of funders or researchers' attachment to the biomedically informed model of evaluation discussed above, or due to an (understandable) preoccupation with developing an evidence base about *what* works over understanding *how* it works, much evaluation of complex safety interventions in healthcare has tended to prioritise (quantitative) demonstration of impact over (qualitative) understanding of process. Yet understanding how interventions work is crucial, for the reasons discussed in the previous section. Patient safety interventions are usually as complex as any, and their reliance on co-ordinated efforts across a variety of actors working at different levels in the system means that consistency of impact can be infuriatingly elusive. Attention to the variety of contextual influences that might mean that a seemingly robust, 'proven' intervention that works in one context might not work in another is a fundamental part of developing an actionable, reliable, transferable evidence base.

One example in particular will suffice to demonstrate this point. The 2000s saw a number of studies that showed seemingly impressive impacts from checklist-based safety interventions in healthcare, including in bloodstream infections in intensive care (Pronovost *et al.* 2006), peri-operative care (De Vries *et al.* 2010), and surgery (Haynes *et al.* 2009). These evaluations were published in high-impact journals (the field-leading *New England Journal of Medicine* in these three cases), deploying relatively robust study designs located at respectable tiers of the hierarchy of evidence (controlled or uncontrolled before-and-after studies—Levels III or IV in Figure 1). Their approach to change had its roots in an intervention (the checklist) that is widely used in other safety-critical fields such as civil aviation, and to which much success has been attributed (Clay-Williams and Colligan 2015). They thus had some promise in theory—and in practice, results seemed to indicate significant impact and potential to improve patient safety worldwide. In the case of the surgical safety checklist (Haynes *et al.* 2009), much was made of the universality of the intervention: it was evaluated in a selection of hospitals with divergent safety records, across high-, low- and middle-income countries. Across this diverse and globally representative sample, the evaluation showed a reduction in rates of complication within 30 days of 36 per cent, and a reduction in 30-day mortality rates of 47 per cent (Haynes *et al.* 2009), both comfortably statistically significant.

Here, then, was the Holy Grail: an intervention that was cheap, relatively easily implemented, and proven across a range of contexts rather than being specific to a particular organisation or a narrowly defined patient group. Correspondingly, the WHO, which had produced the guidance on which it was based, endorsed the checklist and encouraged its adoption worldwide, a plea that was enforced by patient safety bodies in many countries, such as the National Patient Safety Agency in the UK.

Yet subsequent experiences, and accompanying research and evaluation, complicated this picture. In particular, extensive qualitative examination of the implementation of the surgical safety checklist and similar tools in multiple contexts has highlighted the challenges of incorporating it into routine work, ensuring consistent use, and—perhaps most importantly of all—securing a culture in which the importance of such activity is embraced. This research reinforces the point that, as the authors of a systematic review put it (Bergs *et al.* 2015: 781), the checklist and its implementation are "a complex social intervention with an expectation of interaction and cooperation between surgeons, anaesthetists and nurses," not a simple fix with a single, linear, predictable, reproducible causal pathway. Research on other checklist-based interventions has similarly exposed the rich, social mechanisms and contextual conditions that underpin success (Dixon-Woods *et al.* 2011), and in whose absence such interventions can fall far short of their initial promise (Dixon-Woods *et al.* 2013).

Findings of this kind will come as no surprise to social scientists. Indeed, they were to some extent anticipated by the investigators of the surgical safety checklist study itself, who found variation in the scale and nature of change, and noted that "the exact mechanism of improvement is less clear and most likely multifactorial" (Haynes *et al.* 2009: 496–7), including variation in the culture and readiness for change of hospitals and teams. But the desirability of prospective mixed-method evaluation of safety interventions, rather than retrospective analyses that seek to explain what went right in apparently successful trials (or what went wrong when the intervention was incorporated into routine practice), has taken longer to establish, notwithstanding the recommendations of bodies such as the MRC. Where integrated qualitative evaluations have been used prospectively, however, they have offered important insights into both the successes and the failures of efforts to improve healthcare

safety.

One notable example, declared "a model for the field" (Pronovost *et al.* 2011: 341), is the evaluation of the first phase of the Health Foundation-funded Safer Patients Initiative. This large-scale programme involved intervention at multiple levels within acute hospitals to instigate changes that would result in tangible improvements in patient safety, with a principal intended outcome of halving the number of adverse events within two years (Health Foundation 2011). The ambition of the Initiative was matched by its evaluation, which comprised a programme of independent studies, including a mixed-methods controlled before-and-after study that measured change in multiple outcomes, and incorporated an extensive ethnographic examination of the Initiative's realisation in four participating hospitals (Benning *et al.* 2011). The evaluation's findings, though, were disappointing. They suggested statistically significant improvement relative to control hospitals on only one of several measures of safety climate, and one measure of clinical process (monitoring of vital signs). Other measures were unmoved compared to the controls, or even showed trends favouring the control group. The integrated ethnographic study cast light on some of the reasons for this: the Initiative was greeted with enthusiasm by senior managers in the participating hospitals, but failed to penetrate to the level of the ward or operating theatre, where it was seen as an elite preoccupation with limited relevance for day-to-day clinical practice: "somewhere between the blunt end and the sharp end, the model of participative engagement on which [the Initiative] was based had got rather lost" (Benning *et al.* 2011: 9).

While the findings of the evaluation were disappointing, the evaluation itself is noteworthy, and represents an advance on previous study designs, for at least two reasons. One is the use of contemporaneous controls, and the deployment of a 'difference-in-difference' analysis in summatively evaluating effectiveness. Had the Initiative been evaluated using an uncontrolled before-and-after design, it might (probably erroneously) have been declared a success, since there were significant improvements in several safety indicators over time (in both the intervention and control groups)—the so-called 'rising tide' effect (Chen *et al.* 2016). Incorrectly attributing this improvement to the Safer Patients Initiative could plausibly have resulted in the allocation of scarce healthcare resources to an expensive but ineffective intervention. Second, the use of qualitative methods as part of the study design offered explanatory purchase, indicating what went wrong, which aspects of the intervention showed promise, and what kinds of modifications might be made in any revised version. Learning of this kind is of course essential to the endeavour of improving safety; null or negative evaluations are not in vain if they provide a resource for improving future improvement efforts. Such insights are likely to come from well designed qualitative work that provides an explanatory complement to the findings of quantitative evaluation.

Other studies of safety interventions in healthcare have followed this mixed-methods model, combining robust quantitative evaluation that aims for the upper tiers of the hierarchy of evidence with qualitative work that defies its position at the bottom of (some versions of) the hierarchy through its explanatory value. The interventions evaluated have arguably become even more sophisticated, drawing on a range of social scientific insights for their theoretical foundation, and seeking to provide the right kind of impetus, in the right quantity, in the relevant parts of the healthcare system, to instigate improvement (see Table 1). But the improvements yielded have often been modest at best—a point to which we return in the next section of the chapter. The most notable examples tend to come from areas such as surgery and intensive care that present high risks to patient safety, but are relatively contained, and thus

perhaps more amenable to concerted, focused improvement efforts.

[Insert Table 1 about here]

The success of the US-based programme to prevent bloodstream infections in intensive care units briefly mentioned above (Pronovost *et al.* 2006) led to an effort to replicate this work in the UK. This effort, however, failed to match the spectacular reductions (and in some cases eliminations) achieved in the original programme, demonstrated by a controlled study (Bion *et al.* 2013) and explained by an integrated ethnography (Dixon-Woods *et al.* 2013). The ethnography pointed towards the role of a policy and organisational environment that discouraged engagement with the programme, and the absence of the kind of 'social movement' and sense of common endeavour that had been instrumental in making a success of the original.

In surgery, one effort to improve safety sought to deploy, in various combinations, the introduction of standard operating procedures, the principles of Lean process improvement, and the use of teamwork training based on aviation's Crew Resource Management to pursue improvement in operating theatre practices, based on the theoretically and empirically informed hypothesis that interventions seeking to address culture and systems in tandem would be more likely to succeed than narrower interventions alone. A series of controlled before-and-after studies and an integrative quantitative analysis offered some support to this hypothesis, finding greater impacts in combined-intervention sites (McCulloch *et al.* 2016). An accompanying qualitative interview-based study sought to explain the differential impact of these approaches, highlighting in particular the significance of intensive expert support in the more successful sites—a finding with important implications for any effort to replicate or roll out the approach (Flynn *et al.* 2016).

Finally, outside perioperative care, Lawton *et al.* (2017) and Sheard *et al.* (2017) presented results from, respectively, the cluster-randomised controlled trial and process evaluation of the Patient Reporting and Action for a Safe Environment (PRASE) intervention. The result of the trial was 'negative', with no significant difference in the primary outcome between the intervention and control group. The process evaluation, however, demonstrated a wide range of responses to the intervention, and offered rich narratives about its use in participating services. At least according to the qualitative data, then, PRASE had potential. Moreover, the trial relied for its primary outcome measure on a data source—the NHS Safety Thermometer—that the authors themselves acknowledged was a highly flawed way of seeking to gauge the impact of an intervention with much broader improvement ambitions. PRASE sought to empower healthcare staff to focus on improving areas that they saw as a priority, based on the feedback of patients. As Lawton *et al.* (2017: 629) noted, "this makes it difficult to predict in advance what changes a ward will choose to make and therefore what outcomes it might be appropriate to measure." Researchers' choices about how to measure the impact of a complex, multifaceted intervention are therefore hugely significant, and often limited by the availability of routinely collected data. An ill-chosen outcome measure provides a poor yardstick against which to judge an intervention, and finding an appropriate outcome measure is all the more difficult when evaluating complex safety interventions. Correspondingly, the biomedical tendency to place so much emphasis on the achievement of statistically significant improvement against a single primary outcome measure is all the more problematic—a theme to which we return in the final section.

The unfulfilled promise of mixed-methods evaluations?

The increasing methodological sophistication of evaluations of safety interventions in healthcare, then, has in general led to increasing pessimism (or perhaps more accurately, cautious realism) about their impact. As noted above, in itself this is no bad thing. The learning to be derived from rigorous evaluation of unsuccessful interventions is at least as great as the potential learning from success (Peden *et al.* 2019; Stephens *et al.* 2018)—though not always as readily published and communicated to the research and practice community (Shojania and Grimshaw 2005). The science of improvement in healthcare is still in its infancy, and the evidence base for the impact and suitability of different forms of intervention in different circumstances is still incomplete. There is much to be gained from the accumulation of knowledge about the what, the how and the when and where, and from the curation of this evidence base in forms that are accessible and useful to researchers and practitioners (Dixon-Woods and Martin 2016; Webb 2011).

On the other hand, a nagging concern for some is that there is something distinctive about intervening in relation to healthcare quality and safety which requires rethinking—perhaps radically—the nature of the evidence base we demand, and the tools we use to produce it. A longstanding debate within the improvement community has focused on whether the traditional methods of health services research are fit for the purpose of evaluating safety interventions, given their complex nature, the iterative way in which they are applied, and the challenges of applying techniques such as randomisation and blinding to a dynamic field (Leape *et al.* 2002; Shojania 2013; Wachter 2010). The studies discussed in the previous section certainly show that, with funding and tenacity, it is *feasible* to apply high-quality quantitative evaluation methods to safety interventions, and to supplement and enrich these with qualitative methods. But the inconsistency of findings, and the disappointment that often follows investment in promising interventions and their evaluation, does raise the question of whether the notion of a 'proven intervention' is something of a chimera—at least if we use the word 'proven' in the sense traditionally used in health services research. Both the hierarchy of evidence, and the established approach to disseminating and implementing evidence in healthcare, revolve around the notion of 'gold standards' of evidence and practice. In other words, they prize interventions that have survived the most rigorous of evaluation, and are thus seen to be *the (single) best practice*, which should accordingly be applied by all practitioners in all circumstances. Again, one can see how such expectations might reasonably be applied to certain kinds of healthcare intervention, particularly pharmaceutical therapies (though again, there is a danger that this neglects the heterogeneity of patient populations, particularly multimorbid groups, as well as marginalising patient preferences: see, e.g., Greenhalgh *et al.* 2014). Given the crucial role that contextual variability can play in improvement efforts, however—as well as the difficulties in specifying an intervention so tightly that it can be applied, near-identically, by different practitioners with different skill sets in different circumstances—aspiring to develop and evidence 'gold-standard' interventions might be a fool's errand.

A more modest—but more pressing and potentially more fruitful—task might be the development of a better understanding of the menu of improvement approaches that is available, and of the likely prerequisites for their effectiveness. Such an evidence base would acknowledge the fallibility of every approach, and make no claim to offer ready-made solutions. It would rather provide as much detailed and helpful guidance as possible on the measures that might be taken to make the approach work—while acknowledge the importance

of the skill, determination and individual style given to any intervention by those leading it. The importance of different configurations of contexts and mechanisms that can causally drive different outcomes is, of course, explicitly acknowledged by approaches such as realist evaluation (see, e.g., Randell *et al.* 2014). Such approaches account for both the range of causal mechanisms that can give rise to outcomes in different circumstances, and the range of desirable endpoints beyond the principal outcomes that traditional biomedical evaluation tends to prize. There is an important balance to be struck, however, between the search for the 'right' configuration of context, mechanism and outcome, and the idiographic description a thousand possible combinations, each unique to its circumstance (Marchal *et al.* 2012). Equally, the reduction of a 'successful' intervention in healthcare safety to the achievement of a statistically significant result against a single, narrow measure that cannot account for broader benefits is also unhelpful—and risks the premature disposal of complex interventions to which a simple biomedical model of evaluation cannot do justice.

Part of the challenge for evaluation, then, is to be able to say something more than that 'nothing works' (if a truly universal gold standard, according to a particular narrow measure, is the required benchmark) or that 'nearly everything works' (in the right, very particular, circumstances). Both the rich description of improvement practices to demonstrate how, when and why, for example through comparative qualitative case studies, and the evidence provided by robust quantitative evaluation that shows what *can* work, have a role in this. But an equally important task is effective curation of the evidence base, and greater interaction between research and practice communities to assist the task of translating research into practice in a context-sensitive way, that allows practitioners to make their own sense of the evidence base—notwithstanding the variety of challenges involved in such efforts (e.g. Chew *et al.* 2013; Martin *et al.* 2011; Ward *et al.* 2009).

Besides this, debates familiar from other fields where distinct or even conflicting research paradigms have come together, continue to preoccupy methodologists in healthcare safety. Methodological purists suggest that the positivist and constructivist roots of epidemiology and qualitative social science respectively make them irreconcilable (Doyle *et al.* 2009). Others take a more pragmatic approach, valuing the complementary insight that qualitative and quantitative understandings can bring in producing local understanding, without concerning themselves with epistemological neatness. Yet these divergent ways of knowing can have very practical ramifications. Rarely, for example, does it prove possible to use qualitative data to exactingly account for divergent outcomes between cases that, by quantitative measures, appear to be 'successful' or 'unsuccessful'. Qualitative insights tend to be more fluid and less certain. But the expectation that qualitative explanations can be neatly mapped onto divergent quantitative results risks giving rise to formulaic, contrived, even positivistic accounts, with a simplicity as beguiling as it is misleading. In this process, the richness and ambiguity of qualitative analysis is reduced to a series of factors: *x* variables that predict the *y* variable of the quantitative outcome. Good qualitative analysis should not be reduced to a regression analysis. Similarly, an inappropriately positivistic mindset among qualitative researchers can impede the accumulation of useful knowledge. Mitchell *et al.* (2017), for example, note a tendency among evaluators to continue to construct checklist-based interventions as simple, stable interventions with predictable intended causal pathways—rather than as something that interacts with its context to produce multiple effects, both intended and unintended, positive and negative. This results, they argue, in the repeated rediscovery of the complexity context dependency of the checklist, rather than the steady accumulation of insights that might help

improve the intervention.

There are debates, too, within the qualitative healthcare research community about the methods most appropriate to providing explanatory insight into safety interventions. The hierarchy of evidence of the biomedical research tradition has parallels within qualitative research, including the validity of ethnography 'in situ' versus reliance on retrospective interview accounts (Silverman 2001), and the utility of insights provided by observational methods that fall short of full ethnographies in the anthropological tradition in their scale and scope (Cupit *et al.* 2018; Leslie *et al.* 2014; Waring and Jones 2016). More prosaically, as public finances in many countries remain under tight control and research and evaluation funders demand maximum value from their investments, opportunities for high-quality mixed-methods evaluations may become rarer. Qualitative researchers in the field of healthcare must be vigilant that the biomedical tradition does not result in a return to a situation where their contribution is seen as a desirable, but disposable, add-on.

Conclusion

Mixed-methods research and evaluation in healthcare quality and safety offers a useful model for safety research elsewhere, particularly in its increasingly sophisticated efforts to bring together qualitative and quantitative insights. But it also provides some warnings for other researchers, particularly in relation to the need to contain expectations about what mixed-methods approaches can provide, and around the considerations important in deriving the maximum value from both lenses—and reconciling them in a satisfactory way.

We have traced a narrative in healthcare safety research in which an understanding of the need for complex, theory-based interventions has been followed, rather belatedly, by sophistication in evaluation methodology. We have attributed this tardiness in part to the long shadow cast by the biomedical evaluation tradition. Arguably, the translation of the findings of this research base into the kind of useable knowledge that can inform practitioner-led work to improve healthcare safety has been even slower, and here again the tendency to rely on traditional biomedical approaches to knowledge translation may in part be culpable. Sophisticated interventions demand sophisticated evaluation, and sophisticated evaluations demand sophisticated approaches to making findings accessible and relevant. This means overcoming the binary notion that something either works (universally and unequivocally) or does not, and taking seriously the critical, reflexive, intelligent work by practitioners that is needed to make any intervention work. Such approaches to curation and translation remain in their infancy in healthcare.

References

Benning, A., Ghaleb, M., Suokas, A., Dixon-Woods, M., et al. (2011) Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation, *BMJ*. **342**,d195.

Bergs, J., Lambrechts, F., Simons, P., Vlayen, A., et al. (2015) Barriers and facilitators related to the implementation of surgical safety checklists: a systematic review of the qualitative evidence, *BMJ Quality & Safety*. **24**, 12, 776–86.

Bion, J., Richardson, A., Hibbert, P., Beer, J., et al. (2013) 'Matching Michigan': a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England, *BMJ Quality & Safety*. **22**, 2, 110–23.

Boyd, C.M. and Kent, D.M. (2014) Evidence-based medicine and the hard problem of multimorbidity, *Journal of General Internal Medicine*. **29**, 4, 552–3.
Braithwaite, J. (2018) Changing how we think about healthcare improvement, *BMJ*. **361**,k2014.
Brennan, T.A., Leape, L.L., Laird, N.M., Hebert, L., et al. (1991) Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I, *New England Journal of Medicine*. **324**, 6, 370–6.
Campbell, M., Fitzpatrick, R., Haines, A., Kinmonth, A.L., et al. (2000) Framework for design and evaluation of complex interventions to improve health, *BMJ*. **321**, 694–6.
Chen, Y.-F., Hemming, K., Stevens, A.J. and Lilford, R.J. (2016) Secular trends and evaluation of complex interventions: the rising tide phenomenon, *BMJ Qual Saf*. **25**, 5, 303–10.
Chew, S., Armstrong, N. and Martin, G. (2013) Institutionalising knowledge brokering as a sustainable knowledge translation solution in healthcare: how can it work in practice? *Evidence & Policy*. **9**, 3, 335–51.
Clay-Williams, R. and Colligan, L. (2015) Back to basics: checklists in aviation and healthcare, *BMJ Qual Saf*. **24**, 7, 428–31.
Cochrane, A.L. (1972) *Effectiveness and efficiency: random reflections on health services*. London: Nuffield Provincial Hospitals Trust.
Craig, P., Dieppe, P., Macintyre, S., Michie, S., et al. (2008) Developing and evaluating complex interventions: the new Medical Research Council guidance, *BMJ*. **337**,a1655.
Cupit, C., Mackintosh, N. and Armstrong, N. (2018) Using ethnography to study improving healthcare: reflections on the 'ethnographic' label, *BMJ Qual Saf*. **27**, 4, 258–60.
De Vries, E.N., Prins, H.A., Crolla, R.M.P.H., den Outer, A.J., et al. (2010) Effect of a comprehensive surgical safety system on patient outcomes, *New England Journal of Medicine*. **363**, 20, 1928–37.
Department of Health (2000) *An organisation with a memory*. London: The Stationery Office.
Department of Health (2006) *Best research for best health: a new national research strategy*. London: Department of Health.
Dixon-Woods, M., Bosk, C.L., Aveling, E.-L., Goeschel, C.A., et al. (2011) Explaining Michigan: developing an ex post theory of a patient safety program, *Milbank Quarterly*. **89**, 2, 167–205.
Dixon-Woods, M., Leslie, M., Tarrant, C. and Bion, J. (2013) Explaining Matching Michigan: an ethnographic study of a patient safety program, *Implementation Science*. **8**,70.
Dixon-Woods, M. and Martin, G.P. (2016) Does quality improvement improve quality?, *Future Hospital Journal*. **3**, 3, 191–4.
Doyle, L., Brady, A.-M. and Byrne, G. (2009) An overview of mixed methods research, *Journal of Research in Nursing*. **14**, 2, 175–85.
Flynn, L.C., McCulloch, P.G., Morgan, L.J., Robertson, E.R., et al. (2016) The Safer Delivery of Surgical Services Program (S3): explaining its differential effectiveness and exploring implications for improving quality in complex systems, *Annals of Surgery*. **264**, 6, 997–1003.
Greenhalgh, T., Howick, J. and Maskrey, N. (2014) Evidence based medicine: a movement in crisis? *BMJ*. **348**,g3725.
Haynes, A.B., Weiser, T.G., Berry, W.R., Lipsitz, S.R., et al. (2009) A surgical safety checklist to reduce morbidity and mortality in a global population, *New England Journal of Medicine*. **360**, 5, 491–9.

Health Foundation (2011) *Learning report: Safer Patients Initiative*. London: The Health Foundation.
Howick, J.H. (2011) *The philosophy of evidence-based medicine*. Oxford: Wiley.
Institute of Medicine (1999) *To err is human: building a safer health system*. Washington, DC: National Academy Press.
Leape, L.L., Berwick, D.M. and Bates, D.W. (2002) What practices will most improve safety? Evidence-based medicine meets patient safety, *JAMA*. **288**, 4, 501–7.
Leslie, M., Paradis, E., Gropper, M.A., Reeves, S., et al. (2014) Applying ethnography to the study of context in healthcare quality and safety, *BMJ Qual Saf*. **23**, 2, 99–105.
Makary, M.A. and Daniel, M. (2016) Medical error—the third leading cause of death in the US, *BMJ*. **353**,i2139.
Marchal, B., van Belle, S., van Olmen, J., Hoeree, T., et al. (2012) Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research, *Evaluation*. **18**, 2, 192–212.
Martin, G., Currie, G. and Lockett, A. (2011) Prospects for knowledge exchange in health policy and management: institutional and epistemic boundaries, *Journal of Health Services Research & Policy*. **16**, 4, 211–7.
McCulloch, P., Morgan, L., Flynn, L., Rivero-Arias, O., et al. (2016) *Safer delivery of surgical services: a programme of controlled before-and-after intervention studies with pre-planned pooled data analysis*. Programme Grants for Applied Research. Southampton: NIHR Journals Library.
McGlynn, E.A., Asch, S.M., Adams, J., Keesey, J., et al. (2003) The quality of health care delivered to adults in the United States, *New England Journal of Medicine*. **348**, 26, 2635–45.
Melnyk, B.M. and Fineout-Overholt, E. (2011) *Evidence-based practice in nursing and healthcare: a guide to best practice*. Philadelphia, PA: Wolters-Kluwer.
Mitchell, B., Cristancho, S., Nyhof, B.B. and Lingard, L.A. (2017) Mobilising or standing still? A narrative review of Surgical Safety Checklist knowledge as developed in 25 highly cited papers from 2009 to 2016, *BMJ Qual Saf*. **26**, 10, 837–44.
Oakley, A. (2000) *Experiments in knowing: gender and method in the social sciences*. Cambridge: Polity Press.
Pawson, R. and Tilley, N. (1997) *Realistic evaluation*. London: Sage.
PCORI (2018) *PCORI methodology standards*. Washington, DC: Patient-Centered Outcomes Research Institute.
Peden, C.J., Stephens, T., Martin, G., Kahan, B.C., et al. (2019) Effectiveness of a national quality improvement programme to improve survival after emergency abdominal surgery (EPOCH): a stepped-wedge cluster-randomised trial, *The Lancet*. In press.
Pope, C. and Mays, N. (1993) Opening the black box: an encounter in the corridors of health services research, *BMJ*. **306**, 315–8.
Pronovost, P., Needham, D., Berenholtz, S., Sinopoli, D., et al. (2006) An intervention to decrease catheter-related bloodstream infections in the ICU, *New England Journal of Medicine*. **355**, 26, 2725–32.
Pronovost, P.J., Berenholtz, S.M. and Morlock, L.L. (2011) Is quality of care improving in the UK? *BMJ*. **342**,c6646.
Randell, R., Greenhalgh, J., Hindmarsh, J., Dowding, D., et al. (2014) Integration of robotic surgery into routine practice and impacts on communication, collaboration, and decision making: a realist process evaluation protocol, *Implementation Science*. **9**, 1, 52.
Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., et al. (1996) Evidence based medicine: what it is and what it isn't, *BMJ*. **312**, 71–2.

Schuster, M.A., McGlynn, E.A. and Brook, R.H. (2005) How good is the quality of health care in the United States? *Milbank Quarterly*. **83**, 4, 843–95.
Shekelle, P.G., Pronovost, P.J., Wachter, R.M., Taylor, S.L., et al. (2011) Advancing the science of patient safety, *Annals of Internal Medicine*. **154**, 10, 693–6.
Shojania, K.G. (2012) Deaths due to medical error: jumbo jets or just small propeller planes?, *BMJ Quality & Safety*. **21**, 9, 709–12.
Shojania, K.G. (2013) Conventional evaluations of improvement interventions: more trials or just more tribulations? *BMJ Quality & Safety*. **22**, 11, 881–4.
Shojania, K.G. and Dixon-Woods, M. (2017) Estimating deaths due to medical error: the ongoing controversy and why it matters, *BMJ Quality & Safety*. **26**, 5, 423–8.
Shojania, K.G. and Grimshaw, J.M. (2005) Evidence-based quality improvement: the state of the science, *Health Affairs*. **24**, 1, 138–50.
Silverman, D. (2001) *Interpreting qualitative data: methods for analysing talk, text and interaction*. London: Sage.
Stephens, T.J., Peden, C.J., Pearse, R.M., Shaw, S.E., et al. (2018) Improving care at scale: process evaluation of a multi-component quality improvement intervention to reduce mortality after emergency abdominal surgery (EPOCH trial), *Implementation Science*. **13**, 1, 142.
Wachter, R.M. (2010) Patient safety at ten: unmistakable progress, troubling gaps, *Health Affairs*. **29**, 1, 165–73.
Ward, V., House, A. and Hamer, S. (2009) Knowledge brokering: the missing link in the evidence to action chain? *Evidence & Policy*. **5**,267–79.
Waring, J., Allen, D., Braithwaite, J. and Sandall, J. (2016) Healthcare quality and safety: a review of policy, practice and research, *Sociology of Health & Illness*. **38**, 2, 198–215.
Waring, J. and Jones, L. (2016) Maintaining the link between methodology and method in ethnographic health research, *BMJ Qual Saf*. **25**, 7, 556–7.
Webb, D. (2011) Foreword. In *Evidence: Safer Patients Initiative phase two*. London: The Health Foundation. pp. iv–vii.
Weiss, C.H. (1995) Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In Connell, J.P., Kubisch, A.C., Schorr, L.B., and Weiss, C.H. (eds) *New approaches to evaluating community initiatives: concepts, methods, and contexts*. New York: Aspen Institute.
World Health Organization (2018) *10 facts on patient safety*. https://www.who.int/features/factfiles/patient_safety/en/. Accessed 26 April 2019.

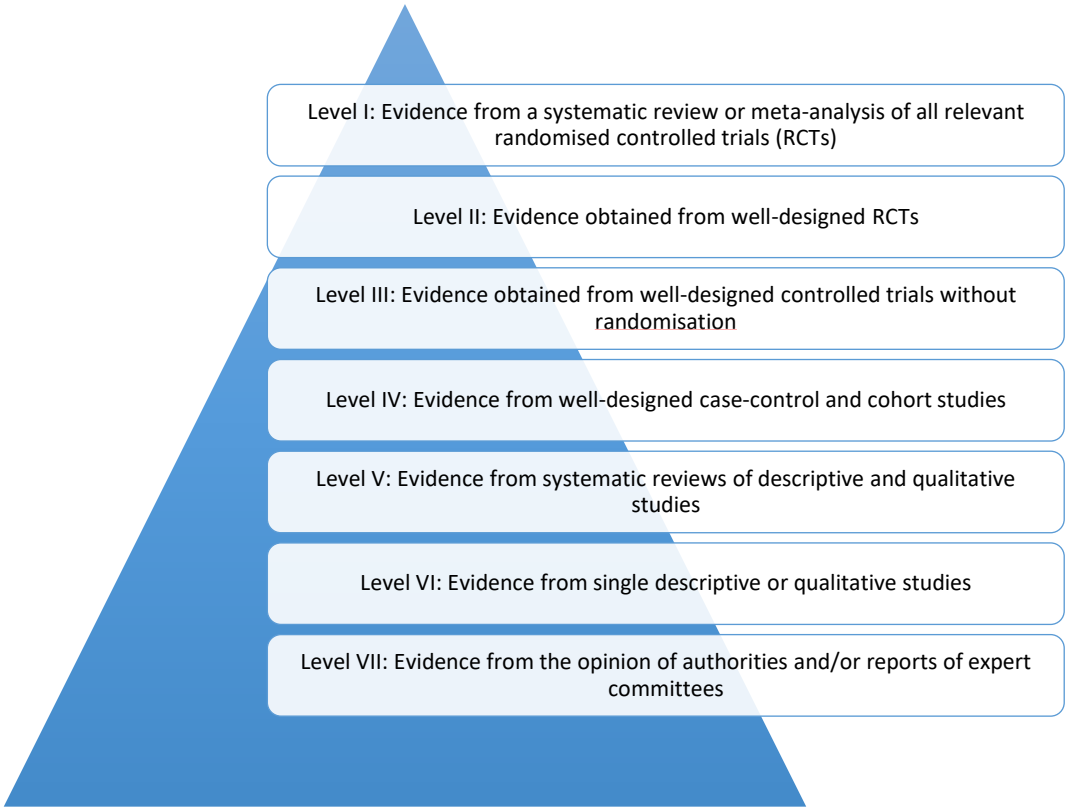

Figure 1: The hierarchy of evidence. Source: Text from Melnyk and Fineout-Overholt (2011: 12)

Programme	Setting	Programme aims and measures	Programme methods	Evaluation approach	Evaluation findings	Further information
Safer Patients Initiative	Hospital wide: wards, critical care, perioperative care, medicines management	Improve organisation-wide safety as measured by a variety of indicators, including reductions in mortality and adverse events	Implementation of evidence-based care bundles in key clinical areas, supported by leadership and cultural interventions at the hospital level	Controlled before-and-after study incorporating difference-in-difference analysis, with integrated ethnographic study	Most measures show no significant difference between control and intervention groups; ethnography suggests limited awareness or impact of work at sharp end	Benning *et al.* (2011)
Matching Michigan	Adult and paediatric intensive care units	Improve practices in intensive care to reduce catheter-associated bloodstream infection rates	Use of technical interventions to ensure consistent use of evidence-based practices known to improve infection control and non-technical interventions to improve systems and culture; introduction of a national catheter-related bloodstream infection reporting system	Non-randomised, stepped study of roll-out of intervention, comparing data before and after implementation, with integrated ethnographic study	The rate of improvement in infection control practices did not significantly change after the intervention; ethnography highlights a hostile policy and management environment and a failure to replicate the social movement of the prototype Michigan approach	Bion *et al.* (2013); Dixon-Woods *et al.* (2013)
Safer Delivery of Surgical Services (3S) programme	Surgery	Improve various aspects of surgical care and outcomes, including teamwork, error rates, patient-reported outcomes, mortality, length of stay, readmissions	Use of three interventions, each designed to address a different cause of patient safety issues: teamwork training based on Crew Resource Management; introduction of standard operating procedures; and process improvement based on Lean principles	Controlled evaluations of combinations of the intervention, with meta-analysis; retrospective qualitative interviews with team and participants	Some indication that combined approaches may address some aspects of surgical safety; interviews shed light on the reasons for differential impact and the importance of strong support from the research team that may not be replicable	McCulloch *et al.* (2016); Flynn *et al.* (2016)
Patient Reporting and Action for a Safer Environment (PRASE)	Ward-based care	Improve safety of care for hospital inpatients, measured by proportion of patients receiving 'harm free care' (Safety Thermometer)	Introduction of action planning cycle: (i) facilitate collection and analysis of patient reported safety concerns; (ii) feed back to staff, building teams' intelligence about patient experience of safety, areas for improvement, and capacity to improve quality.	Cluster-randomised controlled trial; observation of action planning meetings and qualitative interviews with staff	No significant difference on the primary outcome between control and intervention arms; qualitative study suggests enthusiasm for intervention but differential engagement across wards	Lawton *et al.* (2017); Sheard *et al.* (2017)

Table 1: Four complex interventions to improve quality or safety in healthcare, and their evaluations