HEALTH SERVICES AND DELIVERY RESEARCH

VOLUME 4 ISSUE 16 MAY 2016
ISSN 2050-4349

Challenges, solutions and future directions in the evaluation of service innovations in health care and public health

Rosalind Raine, Ray Fitzpatrick, Helen Barratt, Gywn Bevan, Nick Black, Ruth Boaden, Peter Bower, Marion Campbell, Jean-Louis Denis, Kelly Devers, Mary Dixon-Woods, Lesley Fallowfield, Julien Forder, Robbie Foy, Nick Freemantle, Naomi J Fulop, Elizabeth Gibbons, Clare Gillies, Lucy Goulding, Richard Grieve, Jeremy Grimshaw, Emma Howarth, Richard J Lilford, Ruth McDonald, Graham Moore, Laurence Moore, Robin Newhouse, Alicia O'Cathain, Zeynep Or, Chrysanthi Papoutsi, Stephanie Prady, Jo Rycroft-Malone, Jasjeet Sekhon, Simon Turner, Samuel I Watson and Merrick Zwarenstein

NHS
National Institute for
Health Research

DOI 10.3310/hsdr04160

Challenges, solutions and future directions in the evaluation of service innovations in health care and public health

Rosalind Raine,[1]* Ray Fitzpatrick,[2]* Helen Barratt,[3] Gywn Bevan,[4] Nick Black,[5] Ruth Boaden,[6,7] Peter Bower,[8] Marion Campbell,[9] Jean-Louis Denis,[10] Kelly Devers,[11] Mary Dixon-Woods,[12] Lesley Fallowfield,[13] Julien Forder,[14] Robbie Foy,[15] Nick Freemantle,[16] Naomi J Fulop,[1] Elizabeth Gibbons,[2] Clare Gillies,[17] Lucy Goulding,[18] Richard Grieve,[19] Jeremy Grimshaw,[20] Emma Howarth,[21] Richard J Lilford,[22] Ruth McDonald,[6] Graham Moore,[23] Laurence Moore,[24] Robin Newhouse,[25] Alicia O'Cathain,[26] Zeynep Or,[27] Chrysanthi Papoutsi,[28,29] Stephanie Prady,[30] Jo Rycroft-Malone,[31] Jasjeet Sekhon,[32] Simon Turner,[1] Samuel I Watson[22] and Merrick Zwarenstein[33]

[1]Department of Applied Health Research, University College London, London, UK
[2]Health Services Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[3]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames, Department of Applied Health Research, University College London, London, UK
[4]Department of Management, London School of Economics and Political Science, London, UK
[5]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK
[6]Alliance Manchester Business School, University of Manchester, Manchester, UK
[7]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Greater Manchester, Manchester, UK
[8]National Institute for Health Research (NIHR) School for Primary Care Research, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK
[9]Health Services Research Unit, University of Aberdeen, Aberdeen, UK

[10]Canada Research Chair in Governance and Transformation of Health Organizations and Systems, École Nationale d’Administration Publique, Ville de Québec, QC, Canada
[11]Health Policy Centre, Urban Institute, Washington, DC, USA
[12]Department of Health Sciences, University of Leicester, Leicester, UK
[13]Sussex Health Outcomes Research and Education in Cancer (SHORE-C), University of Sussex, Brighton, UK
[14]School of Social Policy, Sociology and Social Research, University of Kent, Canterbury, UK
[15]Academic Unit of Primary Care, Leeds Institute of Health Sciences, Faculty of Medicine and Health, University of Leeds, Leeds, UK
[16]Department of Primary Care and Population Health, University College London, London, UK
[17]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) East Midlands and NIHR Research Design Service East Midlands, University of Leicester, Leicester, UK
[18]King’s Improvement Science, Centre for Implementation Science, King’s College London, London, UK
[19]Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK
[20]Clinical Epidemiology Program, Ottawa Hospital Research Institute and Department of Medicine, University of Ottawa, Ottawa, ON, Canada
[21]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) East of England, University of Cambridge, Cambridge, UK
[22]Warwick Medical School, University of Warwick, Coventry, UK
[23]School of Social Sciences, Cardiff University, Cardiff, UK
[24]Medical Research Council (MRC)/Chief Scientist Office (CSO) Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK
[25]Indiana University School of Nursing, Indianapolis, IN, USA
[26]School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[27]Institut de Recherche et Documentation en Économie de la Santé, Paris, France
[28]Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[29]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Northwest London, Imperial College London, London, UK
[30]Department of Health Sciences, University of York, York, UK
[31]School of Healthcare Sciences, Bangor University, Bangor, UK
[32]Department of Political Science and Statistics, University of California Berkeley, Berkeley, CA, USA
[33]Centre for Studies in Family Medicine, Department of Family Medicine, Western University, London, ON, Canada

Declared competing interests of authors: none

*Corresponding author

Published May 2016

DOI: 10.3310/hsdr04160

This report should be referenced as follows:

Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16).

Health Services and Delivery Research

ISSN 2050-4349 (Print)

ISSN 2050-4357 (Online)

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: nihredit@southampton.ac.uk

The full HS&DR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hsdr. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

Criteria for inclusion in the *Health Services and Delivery Research* journal

Reports are published in *Health Services and Delivery Research* (HS&DR) if (1) they have resulted from work for the HS&DR programme or programmes which preceded the HS&DR programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

HS&DR programme

The Health Services and Delivery Research (HS&DR) programme, part of the National Institute for Health Research (NIHR), was established to fund a broad range of research. It combines the strengths and contributions of two previous NIHR research programmes: the Health Services Research (HSR) programme and the Service Delivery and Organisation (SDO) programme, which were merged in January 2012.

The HS&DR programme aims to produce rigorous and relevant evidence on the quality, access and organisation of health services including costs and outcomes, as well as research on implementation. The programme will enhance the strategic focus on research that matters to the NHS and is keen to support ambitious evaluative research to improve health services.

For more information about the HS&DR programme please visit the website: http://www.nets.nihr.ac.uk/programmes/hsdr

This report

The report in this issue of the journal is the output from a meeting in London in June 2015 and the subsequent management of the proceedings funded by The Health Foundation, The Medical Research Council, the National Institute for Health Research and Universities UK. The final report began editorial review in November 2015 and was accepted for publication in March 2016. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HS&DR editors and production house have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the final report document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent work and the views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HS&DR programme or the Department of Health.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).

Contents

Foreword

Providing health care is among the most complex and risky activities carried out in the world. Each person and their circumstances are unique; care is delivered by frontline staff often working in teams and in many different settings, supported by teams of managerial and administrative staff, and using complex technologies. The outcomes for patients that materialise depend on a multitude of factors that relate to the context of care delivery; for example, the quality of the interactions between staff and patient, established practice, national and local initiatives designed to improve care, and the mix of financial and non-financial incentives at play. The task of the evaluator hoping to assess accurately the impact of a programme or initiative is to understand how interventions interact with this surrounding context. As many have stated before, 'intervention plus context equals outcome'.[1] And, the more complex an intervention is, the more likely context is to influence outcome.

This volume brings together contributions of leading thinkers to present a 'state of the art' in the evaluation of complex interventions. It is clear from the contributions in many essays that the task of evaluating complex interventions is almost as complex as the systems being evaluated. Important questions include how can evaluators describe interventions and context and the relationship between these? How can observational studies deal with internal validity, when individuals are recruited into treatments in such heterogeneous ways? What is the role of randomised controlled trials in situations where health-care practitioners and service users may be far from equipoise? How can the evaluation findings be synthesised when so much depends on the detail of local implementation and context?

These are technical questions, but the need now to renew thinking has never been clearer. In England, the NHS faces an unprecedented financial squeeze, adult social care faces significant cuts, and there is huge pressure on investment in prevention and population health. This profound challenge is coupled with rising numbers of people living with multiple and long-term comorbidities and needing care and support. These pressures are common to all countries, although they may materialise differently. They require a complex response and a step change in the historic performance of the health, health care and social care system. The ability of health systems to respond to these challenges depends on skilful design and robust implementation of a range of initiatives, not just once but in a dynamic stream, rooted in and modified by information on impact: in short, rooted in intelligent evaluation that is sensitive to the complexity.

Published May 2016
DOI: 10.3310/hsdr04160-xi

This foreword should be referenced as follows:

Dixon J. Foreword. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al*. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. xi–xii.

We have become aware of the importance of context as well as the intervention in making change: in a masterly review on the subject, Penelope Hawe noted the 'background becomes the foreground'.[2] But health services research has not kept up with this insight: many evaluations appear to be 'acontextual', many just assume that the intervention is as unchanging as a 'pill' in its scope, size and intensity, and many are summative rather than formative. It is timely to reflect on these issues and to reassess the methods used by health services researchers in this area and their contribution to change. This volume is one step in that direction.

Jennifer Dixon

Jennifer Dixon, Chief Executive, The Health Foundation

References

1. Bate P, Robert G, Fulop N, Øvretveit J, Dixon-Woods M. *Perspectives on Context*. London: The Health Foundation; 2014.
2. Hawe P. Lessons from complex interventions to improve health. *Ann Rev Public Health* 2015;**36**:307–23. http://dx.doi.org/10.1146/annurev-publhealth-031912-114421

Foreword

The Medical Research Council's (MRC's) strategy is built around three simple words: we **prioritise**, **discover** and **transform**, so that medical research can change lives.

The MRC supports robust, reproducible advances in medical knowledge, and key to this is the development and implementation of rigorous methodologies for research. As this volume clearly articulates, we are in a time of significant transformation and it is crucial that we take this opportunity to conduct appropriately designed evaluative research, to ensure that health care is delivered in the most evidence-based and effective manner possible.

Partnership is crucial; the MRC and the National Institute for Health Research (NIHR) work hand in hand, through the MRC–NIHR Methodology Research Programme, to develop and implement modern methodologies in biomedical and health-care research. In supporting the development of this volume, we have teamed up with a third partner, The Health Foundation.

No innovation can be said to have been successful without robust evaluation of its effects and – arguably – no innovation should be implemented without plans for evaluation. This volume will support researchers, to ensure that they understand the range and applicability of methods available for the evaluation of system-level innovations in health care and public health. This will enable the generation of the highest quality evidence to support, in turn, the decisions of commissioners of health care and public health services, and of policy-makers.

The excellent work done by the editors and authors of this volume, as well as the participants in the workshop which informed it, should enable a significant enhancement in the quality and quantity of evaluative research around innovations in health care and public health systems.

Much as the MRC, NIHR and The Health Foundation have come together to support this initiative, each bringing unique skills, experience and knowledge, so we encourage researchers, health-care providers and policy-makers to come together to enable a new generation of evaluated health-care innovations. This volume will enable these groups to discover the most up-to-date methodologies for evaluative research, to prioritise these methods and implement them in the most appropriate manner (meeting the needs of both the research community and care providers), and thereby to transform the provision of health care and public health.

John Savill

John Savill, Chief Executive, Medical Research Council

Published May 2016
DOI: 10.3310/hsdr04160-xiii

This foreword should be referenced as follows:

Savill J. Foreword. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al*. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. xiii–xiv.

Foreword

Health-care systems face major challenges. These challenges partly reflect major successes, in particular steadily increasing life expectancy. The health needs of an ageing population are more complex. Challenges also arise from expanding expectations of what might be provided by services as well as the inherent dynamism and inventiveness of medical technology. Challenges also provide opportunities. In the case of health services, policy statements such as NHS England's *Five Year Forward View* make cogent cases that current pressures on services require innovation and experiment to produce new sustainable solutions to health care. This in turn requires research so that innovations are rapidly identified and their potential is evaluated and disseminated.

I am delighted that the resources and distinct expertise of the National Institute for Health Research (NIHR), The Health Foundation, the Medical Research Council and Universities UK have combined to produce this collection of position papers on how to evaluate health care and public health services. Key contributions to the volume came from two parts of NIHR. The Collaborations for Leadership in Applied Health Research and Care (CLAHRCs) are partnerships between universities, the NHS, public health and other relevant organisations to carry out world-class applied health research, particularly in relation to chronic disease and public health interventions. The Health Services and Delivery Research (HSDR) programme aims to produce rigorous and relevant evidence to improve the quality, accessibility and organisation of health services.

The volume provides a clear and authoritative explanation for the range of methods that can now be brought to bear to evaluate services. A wide spectrum of methods are described from novel forms of randomised trials to innovative statistical techniques for analysing data about services, outcome measures focused on patients' priorities, and new focuses of research such as how to implement best practice. An impressive range of experts were mobilised to contribute to the debates out of which the position papers emerged. As well as providing accessible state-of-the-art explanations of best methods for evaluative research, the volume contains other important messages. These messages are that evaluation involves partnership between health professionals, providers, commissioners and researchers; and that innovation will best emerge from early and close dialogue between these different partners.

I congratulate those involved in producing this volume in providing a promising route map for generating evidence to inform the improvement of services.

Sally C

Sally Davies, Chief Medical Officer, Department of Health

Published May 2016
DOI: 10.3310/hsdr04160-xv

This foreword should be referenced as follows:

Davies S. Foreword. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al*. Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. xv–xvi.

Introduction

Ray Fitzpatrick[1] and Rosalind Raine[2]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, UK
[2]Department of Applied Health Research, University College London, London, UK

Declared competing interests of authors: none

Published May 2016
DOI: 10.3310/hsdr04160-xvii

This introduction should be referenced as follows:

Fitzpatrick R, Raine R. Introduction. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. xvii–xxiv.

List of abbreviations

CLAHRC	Collaboration for Leadership in Applied Health Research and Care
HSDR	Health Services and Delivery Research
MRC	Medical Research Council
NETSCC	NIHR Evaluation Trials and Studies Coordinating Centre
NIHR	National Institute for Health Research
PROM	patient-reported outcome measure
RCT	randomised controlled trial

We are all familiar with the dynamic whereby proposed innovations in medicines have to be subjected to detailed scrutiny of evidence claiming to support new products. There is considerable agreement about the types of evidence required, and regulatory arrangements are in place to ensure that such evidence is made available and is as supportive as industry claims. No such consensus exists about the nature and importance of evidence required when changes are proposed to health, social care and public health services. This is a paradox when one considers the claims made for proposed changes in services and the scale of costs and benefits that may follow from changes. This collection of essays is intended to address at least the first part of the paradox, the relative lack of consensus about how to evaluate changes in services.

Concerns about the pressures that face current and future health and social care budgets and that make change urgent are well rehearsed.[1,2] Populations in high-income countries such as the UK have enjoyed steady and consistent improvements in life expectancy. Accompanying these developments have been the increasing numbers of individuals with long-term conditions and the frailties of older age. Among individuals with long-term conditions it is now common for more than one long-term condition to occur.[3] The needs for interventions and support of individuals become more complex and extensive, especially given the scope for modifying the relationship between ageing and disability.[4]

The majority of health-care use, and hence of costs, stems from individuals with long-term conditions and disabilities of ageing. Systems and services need to make multiple, extensive adjustments in response to the complexity and volume of problems presented. It might be argued that, until recently, the default or assumed model of illness around which Western health care has developed has been the single acute problem and its diagnosis and management; that is no longer appropriate for the complex and long-term medical, social and emotional features of modern morbidity. It is now recognised that heterogeneity of treatment effects and the organisation and delivery of the same intervention, together with the need to shift from the perspective of the service to that of the patient with respect to care delivery across diseases and settings, must all be taken into account when evaluating outcomes.

Health services in high-income economies have a fairly consistent tendency to grow in scale, complexity and cost. This partly reflects the changing nature of morbidity outlined above. Individuals with health problems such as diabetes, multiple sclerosis or stroke may receive a wide range of interventions from diverse staff and services; co-ordinating such diversity, along prevention, treatment, rehabilitation and palliation pathways, becomes the challenge. Growing scale and costs are also shaped by the inherent inventiveness and dynamism of health-care sciences. Some analyses identify medical technology as the largest source of growth in costs.[5] Growing societal expectations are also important. By far the most striking example of these trends is found in the USA, which devotes > 17% of Gross Domestic Product to health care (that is exclusive of social care expenditure). No other country nearly matches that proportion of expenditure but all systems show inherent growth that seldom reverses. In summary, the commonest perception of these universal trends is that escalating costs are unsustainable and solutions to manage costs are urgently needed from within health-care systems.

Another striking feature of health care in the USA is its great diversity.[6] Because there are so many funders and providers of services, it is often referred to as pluralist in nature. There are few overarching mechanisms to standardise what services are provided or how they are organised. A common assumption is that systems such as the NHS will have less diversity because of the dominant influence of public funds and centrally determined health policy, plans and specific targets. In reality, and despite the existence of a variety of mechanisms that might standardise services, variations also exist in systems such as the NHS.[7] The very diversity of ways of organising services forcefully raises the question of how best to provide services. This in turn indicates the need for evidence to explain variations and inform decisions to change.

Over 40 years ago, a compelling case was made that, while in many respects an outstanding institution, the NHS had one fundamental weakness, namely a lack of evidence of what works. Cochrane expressed the problem in terms of a complete absence of analysis of relationships between inputs and outputs.[8] The problems that he identified with the evidence base are still relevant to health-care systems today. For example, he raised the issue of how best to capture the outputs of services. He questioned the lack of evidence for the commonest practices in terms of diagnostic tests, length of hospital stay and where and how services are delivered. His book is most often cited for its successful advocacy of the experiment, the randomised controlled trial (RCT), to provide evidence for the health service. Subsequently, the trial became recognised as the most unbiased method for evaluating interventions and in terms of influencing decisions the systematic review of trials is commonly cited as the gold standard.

The RCT is, however, a method that has required adaptation to be useful in the evaluation of systems and services as opposed to drugs or other simple discrete interventions. In 2000, the Medical Research Council (MRC) published guidance for the evaluation of complex interventions, with a view to developing methods for evaluating services which comprise complex and varying elements and contributions, for example a stroke unit, a decision aid or a health promotion campaign.[9] The guidance made the case for the importance of other methods, for example modelling, simulation and qualitative methods, to complement well-established core methods, including the RCT and cohort study. Subsequent versions of the guidance extended the first version and also defined methods for process evaluation on the basis that evidence on how to implement a novel intervention was as important as evidence of its efficacy.[10,11] Other collections of guidance, largely based on NHS Research and Development- and subsequently, National Institute for Health Research (NIHR)-funded work, continued to be published, expanding the range of methods for service evaluation, particularly in areas such as qualitative methods, economic evaluation and evidence synthesis.[12,13]

In the recent past, The Health Foundation has also played a key role in sponsoring work that, among other things, developed methods for evaluation and quality improvement emphasising non-linear, adaptive processes whereby complex health systems change.[14] Many key concepts and insights emerged. Their work emphasised the many ways and levels at which context shapes service development.[15] Such insights highlight the value of shifting from the traditionally used binary question of effectiveness (yes/no), towards a more sophisticated exploration of, for example, generalisable determinants of beneficial outcomes. The Health Foundation supported thinking about the value of early formative assessment where researchers breach conventional independence from services being evaluated to share researchers' emerging insights.[16]

Cochrane himself, and much of the subsequent guidance, acknowledged that many uncertainties about service development, innovation and reorganisation could not be addressed by the classic RCT, but still effectively gave this method primacy of attention and recognition. We have reached the point now where attention in terms of articulating, refining and developing principles can be given to a much wider array of methods. Funders of evaluation have become more pragmatic in supporting research methods fit to address any given problems. Users of evaluative research can gain understanding of problems and potential solutions through many other kinds of evidence in addition to the classic large-scale multicentre RCT. This volume aims to set out the current state of play in a broad menu of methods as well as identifying challenges and potential future directions.

All involved in this volume felt that the need for an authoritative overview was pressing. A key trigger was the publication in 2014 by NHS England of the NHS *Five Year Forward View*.[17] It highlighted the now familiar challenges facing health-care systems in terms of rapidly changing patterns of ill-health and the need for matching service reform. However, very striking and novel were the document's articulation of the urgency of such change, and the suggestion that reconfiguration of services can be as beneficial as the introduction of new technologies, devices and drugs. Notable also was the aspiration that such changes should no longer be centrally driven. Instead, the NHS was encouraged to pursue local and regional experiments to bring about a range of new models of care schematically outlined in the document. Innovation was henceforth to be achieved from the 'bottom up'.[17] Evaluative research was explicitly recommended as part of the solution, but no real detail was provided.

The 'forward view' provided exactly the right stimulus to collaborate for three major UK health research funders whose interests and roles in developing the methodological underpinnings for evaluative research have already been mentioned: The Health Foundation, the MRC and NIHR [specifically via its Health Services and Delivery Research programme and the Collaborations for Leadership in Applied Health Research and Care (CLAHRCs)]. CLAHRCs represent the NIHR infrastructure most directly charged to carry out applied health research in partnership between the NHS and researchers. The common thread to all partners was the sense that bottom-up transformation has to be shown to work, using a range of appropriate research methods, that we need to clarify underlying mechanisms and accommodating contexts and that we must provide research evidence in a timely fashion to ensure effective dissemination and adoption of optimal system and service change.

The three research funding bodies, together with Universities UK, therefore, funded a high-level meeting in London in June 2015 to present and discuss the eight key evaluation themes presented in this volume. Over 90 senior applied health researchers came together for an intense 2 days of challenge and debate. Each of the eight topics was the focus of plenary sessions, with the invited attendees (listed below) discussing the presentations at dedicated sessions. Discussions were recorded and transcripts made available together with plenary presentations to the authors invited to lead in drafting the essays. Many of the lead authors, all experts in the relevant field, were drawn from the 13 CLAHRCs. Drafts were sent to all plenary speakers for critique and revision; revised versions were sent to the two editors, who suggested final modifications. The outcome is the essays collected here, further revised in the light of a set of reviews of independent referees. Prepress Projects Ltd, who support the NIHR Journals Library, provided professional editorial input.

The contributors to this volume responded to the challenge to state as clearly as possible the state of play in the eight chosen areas and to identify remaining challenges. They were given discretion as to how best to express the degree of consensus in their area. As a result, a minority of essays' authors reported the range of views expressed at the London meeting, whereas a majority preferred to use a more conventional format of authorial consensus. The editors, and indeed all involved in producing this volume, recognised that different approaches would be appropriate for different topics. In particular, it was accepted that the degree of consensus would necessarily differ across contributions. This diversity was accepted and welcomed from the outset and is reflected in the range of formats in which different fields are discussed in the contributions.

Samuel I Watson and colleagues provide an excellent starting point by emphasising the central role of evidence in helping decision-makers. They also make clear claims for the importance of models and theories in pulling together how different components of a system may interact. For many problems, new primary research may not be needed and decision-makers can be advised by a synthesis of available evidence. This synthesis has to factor in the quality of different studies. The authors' approach is a Bayesian solution to the various challenges of drawing conclusions from disparate existing evidence.

Helen Barratt and colleagues provide an overview of how the RCT has evolved to be relevant to complex health service/public health research needs. Cluster RCTs and stepped-wedge designs permit randomisation of the unit or organisational entity rather than the individual patient and are, therefore, especially relevant here; the authors clearly delineate their respective advantages and disadvantages. Overall, they argue for pragmatic trial designs developed with service providers, with process evaluation also given strong emphasis.

Quantitative methods relevant to RCTs are now relatively well understood and best practices are agreed. Health services and public health research have a rich and growing body of observational data, largely from NHS records but also from specifically collected quantitative data. Clare Gilles and colleagues provide an authoritative overview of analytic strategies and quantitative methods that may be used to test for causal connections and address potential biases in observational data. Many of the methods have been developed and applied in economics-led data analysis and, to date, less widely used in evaluative health research. They begin by making the strong case for observational data to test the external validity of inferences made in RCTS. They go on to describe a number of methods that will become increasingly widely adopted: difference in difference methods, synthetic controls, instrumental variables and propensity scores.

Patient-reported outcome measures (PROMs) are a potentially powerful method of addressing a key issue in evaluative research. Elizabeth Gibbons and colleagues describe the increasing areas of agreement about how to capture what matters to patients in terms of outcomes. The use of PROMs has steadily expanded from secondary outcomes in clinical trials to main outcomes in major initiatives, such as the NHS's national PROMs programme. To be more widely applicable, for example in evaluating services for long-term conditions, the patient and the health professional need to be engaged to participate in PROMs completion and use in decisions by seeing their potential relevance to routine care.

Rosalind Raine and colleagues provide a cogent reminder of the critical importance of taking a population perspective when evaluating services. They remind us that inequalities of access and outcomes are a key dimension in terms of which to evaluate the performance of services. They show how, to properly address inequalities, researchers have to take account of the different level of health need across the socioeconomic gradient and among different sociodemographic groups in a population. They consider critical issues in designing interventions to address inequalities in health and conclude that interventions must be targeted at multiple levels to work. They also raise population- and system-level issues that may arise from austerity-driven reductions in health-care coverage.

Simon Turner and colleagues address the most challenging of all kinds of evaluation, major system change, for example where broad top-down changes are introduced into a health-care system, as is increasingly happening both regionally and nationally across the NHS. There are numerous difficulties, including the inherent problems of causal narrative with multiple changes at different levels and the influence of the overall social, economic and political context. They provide an overview of examples and, drawing on management and organisational literature, argue for the centrality of theory and of mixed methods to navigate towards explanation of system-level changes in services.

There are echoes of the essay on system changes in the following essay by Emma Howarth and colleagues, who address contextual issues and qualitative research. Qualitative research plays a crucial role in conjunction with quantitative evaluative research, not least in elucidating the specific context in which a trial or evaluative study was carried out. Qualitative research may be used before a trial to develop an intervention, or during or after the trial to assist interpretation and interpretation of results. The authors recognise that research funding is on a trajectory of growing acceptance of qualitative methods. They argue that greater weight and resource should be given to the qualitative component of mixed-methods research studies where investigators have greater appreciation of the need to inform and influence subsequent implementation.

Chrysanthi Papoutsi and colleagues bring the volume to a close with an essay on implementation science, the study of implementing evidence-based practice. The field, although vital, is a relatively new component of evaluative research, so that contributors to this essay were reluctant to be too tied down to specific definitions, approaches or methods, and they make the case that the essay will itself stimulate work towards consensus. As with some other essays, there is agreement on the importance of theory, and a number of specific theoretical options such as process normalisation theory are outlined for the reader.

The authors of the essays were encouraged to think about the future of their chosen subjects and some commonly expressed general points seem to emerge that are optimistic for the future development of evaluative health research and its impact. First and foremost, there is significant consensus on the value of a repertoire of quantitative and qualitative methods working in conjunction to produce relevant evidence. Second, there is shared ground on the importance of researchers working more closely with providers, commissioners and health professionals from the earliest stage to define uncertainties requiring research and to communicate insights and results in suitable formats at different stages. Third, there is an emerging sense of pragmatism, not in the sense of supporting anything other than excellence, but in carrying out research good enough to make timely contributions to inform decisions. Fourth, there is optimism that health care and public health systems will develop in ways that produce increased appetite for research as there is growing shared understanding of its value. Finally, specific features of health care/public health will eventually make research easier, as routine health data, with appropriate safeguards, become more open to use by research and information technology facilitates the integration of health data for the purposes of both care and research.

The eight chosen themes do not define the entire field of evaluative research in this area. They were considered sufficiently important to start a process of defining the state of play of the field but it is recognised that other subjects would undoubtedly emerge in subsequent iterations of the volume. It is also important to express here the hope that the volume itself will be live and open to revision. The NIHR Journals version will have a conventional published form, as with all NIHR monographs, but other platforms will allow commentary and update. It is thereby hoped that this volume will be a continuing stimulus to research that informs the improvement of services and our health, social and public health systems.

Acknowledgements

For their financial support for the meeting in London in 2015 and the subsequent management of the proceedings, the following are very gratefully acknowledged and thanked: The Health Foundation, the MRC and NIHR.

For their organisation of the London meeting and subsequent management of the proceedings, resulting in this publication, Christina Heap (of the NIHR CLAHRC Partnership Programme hosted at Universities UK) and John de Pury (of Universities UK) are also very gratefully acknowledged and thanked.

A small advisory group also gave invaluable strategic thought and steers to the project throughout its development; in addition to John de Pury and Christina Heap, there was Nick Barber (The Health Foundation), David Crosby (MRC), Steph Garfield-Birkbeck [NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC), NIHR Health Services and Delivery Research (HSDR)] and Tara Lamont (NIHR Dissemination Centre and HSDR programme). Their generous effort and consistent support are very gratefully acknowledged.

The following attended and participated in the discussions of the London meeting and their contributions are very gratefully acknowledged:

Janice Baird, University of Southampton; Nick Barber, The Health Foundation; Helen Barratt, University College London; Gwyn Bevan, London School of Economics; Nick Black, London School of Hygiene and Tropical Medicine; Ruth Boaden, Manchester University; Chris Bonell, London School of Hygiene and Tropical Medicine; Peter Bower, University of Manchester; Peter Brocklehurst, University College London; Chris Butler, University of Oxford; David Byrne, Durham University; Marion Campbell, University of Aberdeen; Mike Clarke, Queen's University Belfast; Peter Craig, University of Glasgow; David Crosby, MRC; Steve Cummins, London School of Hygiene and Tropical Medicine; Kelly Devers, Urban Institute, Washington, DC, USA; Jennifer Dixon, The Health Foundation; Mary Dixon Woods, University of Leicester; Jenny Donovan, University of Bristol; Vikki Entwistle, University of Aberdeen; Lesley Fallowfield, University of Sussex; Karen Feinstein, Jewish Healthcare Foundation; Julien Forder, London School of Economics and University of Kent; Robbie Foy, University of Leeds; Nick Freemantle, University College London; Naomi J Fulop, University College London; Mark Gabbay, University of Liverpool; Steph Garfield-Birkbeck, NIHR, NETSCC – HSDR programme; Simon Gates, University of Warwick; John Geddes, University of Oxford; Elizabeth Gibbons, University of Oxford; Clare Gillies, University of Leicester; Lucy Goulding, King's College London; Richard Grieve, London School of Hygiene and Tropical Medicine; Jeremy Grimshaw, University of Ottawa, ON, Canada; Martin Gulliford, King's College London; Julian Higgins, University of Bristol; Richard Hobbs, University of Oxford; Emma Howarth, NIHR CLAHRC East of England; Russ Jago, University of Bristol; David Jones, University of Leicester; Peter Jones, University of Cambridge; Tom Kenny, University of Southampton; Kamlesh Khunti, University of Leicester; Tara Lamont, NIHR NETSCC; Alastair Leyland, University of Glasgow; Richard J Lilford, University of Warwick; Stuart Logan, University of Exeter; Jean-Louis Denis, Montreal, QC, Canada; Suzanne Mason, University of Sheffield; Sue Mawson, University of Sheffield; Nicholas Mays, London School of Hygiene and Tropical Medicine; Ruth McDonald, University of Manchester; Brian Mittman, VA Centre for Implementation Practice and Research Support; Laurence Moore, University of Glasgow; Graham Moore, Cardiff University; Steve Morris, University College London; Bhash Naidoo, National Institute for Health and Care Excellence; Robin Newhouse, University of Maryland, MD, USA; Jon Nicholl, University of Sheffield; Alicia O'Cathain, University of Sheffield; Zeynep Or, Institut de Recherche et Documentation en Economie de la Santé, Paris, France; Chrysanthi Papoutsi, Imperial College London; Mark Petticrew, London School of Hygiene and Tropical Medicine; Kate Pickett, University of York; Terri Piggott, Loyola University of Chicago; Stephanie Prady, University of York; Toby Prevost, King's College London; Julie Reed, NIHR CLAHRC North West London; David Richards, University of Exeter; Anne Rogers, University of Southampton; Jo Rycroft-Malone, Bangor University; Chris Salisbury, University of Bristol; Jane Sandall, King's College London; Jasjeet Sekhon, University of California Berkeley, CA, USA; Nick Sevdalis, King's College London; Lisa Simpson, AcademyHealth; Adam Steventon, The Health Foundation; Matt Sutton, University of Manchester; James Thomas, University College London; Simon Turner, University College London; Kieran Walshe, University of Manchester; Justin Waring, University of Nottingham; Samuel I Watson, University of Warwick; Paula Williamson, University of Liverpool; Paul Wilson, Manchester Business School; and Merrick Zwarenstein, University of Toronto, ON, Canada.

References

1. Astolfi R, Lorenzoni L, Oderkirk J. Informing policy makers about future health spending: a comparative analysis of forecasting methods in OECD countries. *Health Policy* 2012;**107**:1–10. http://dx.doi.org/10.1016/j.healthpol.2012.05.001
2. Appleby J. *Spending on Health and Social Care Over the Next 50 Years: Why Think Long Term?* London: The King's Fund; 2013.
3. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;**380**:37–43. http://dx.doi.org/10.1016/S0140-6736(12)60240-2
4. Christensen K, Doblhammer G, Rau R, Vaupel J. Ageing populations: the challenges ahead. *Lancet* 2009;**374**:1196–208. http://dx.doi.org/10.1016/S0140-6736(09)61460-4
5. Smith S, Newhouse JP, Freeland MS. Income, insurance, and technology: why does health spending outpace economic growth? *Health Aff (Millwood)* 2009;**28**:1276–84. http://dx.doi.org/10.1377/hlthaff.28.5.1276
6. Newhouse J, Garber A, Graham R, McCoy M, Mancher M, Kibria A. *Variation in Health Care Spending: Target Decision Making, Not Geography*. Washington, DC: Institute of Medicine, The National Academies Press; 2013.
7. Public Health England. *The NHS Atlas of Variation in Healthcare*. London: Public Health England; 2015. URL: www.rightcare.nhs.uk/atlas/RC_nhsAtlas3_HIGH_150915.pdf (accessed February 2016).
8. Cochrane A. *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust; 1972.
9. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, *et al.* Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;**321**:694–6. http://dx.doi.org/10.1136/bmj.321.7262.694
10. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;**337**:a1655.
11. Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W. Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 2015;**350**:h1258. http://dx.doi.org/10.1136/bmj.h1258
12. Black N, Brazier J, Fitzpatrick R, Reeves B. *Health Services Research Methods*. London: British Medical Journal Books; 1998.
13. Stevens A, Abrams K, Brazier J, Fitzpatrick R, Lilford R. *The Advanced Handbook in Evidence Based Healthcare*. London: Sage; 2001. http://dx.doi.org/10.4135/9781848608344
14. The Health Foundation. *Complex Adaptive Systems*. London: The Health Foundation; 2010. URL: www.health.org.uk/sites/default/files/ComplexAdaptiveSystems.pdf (accessed February 2016).
15. Bate P, Robert G, Fulop N, Ovretviet J, Dixon-Woods M. *Perspectives on Context*. London: The Health Foundation; 2014. URL: www.health.org.uk/sites/default/files/PerspectivesOnContext_fullversion.pdf (accessed February 2016).
16. The Health Foundation. *Evaluation: What to Consider*. London: The Health Foundation; 2014. URL: www.health.org.uk/sites/default/files/EvaluationWhatToConsider.pdf (accessed February 2016).
17. NHS England, Care Quality Commission, Health Education England, Monitor, Public Health England, Trust Development Authority. NHS: *Five Year Forward View*. London: NHS England; 2014. URL: www.england.nhs.uk/ourwork/futurenhs/2014 (accessed 8 April 2016).

Essay 1 Integrating multiple sources of evidence: a Bayesian perspective

Samuel I Watson and Richard J Lilford

Warwick Medical School, University of Warwick, Coventry, UK

Declared competing interests of authors: Samuel I Watson and Richard J Lilford are both part-funded/supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West Midlands. This essay presents independent research and the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Published May 2016
DOI: 10.3310/hsdr04160-1

This essay should be referenced as follows:

Watson SI, Lilford RJ. Integrating multiple sources of evidence: a Bayesian perspective. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 1–18.

List of figures

List of abbreviations

CPOE	computerised physician order entry
RCT	randomised controlled trial

Abstract

Policies and interventions in the health-care system may have a wide range of effects on multiple patient outcomes and operate through many clinical processes. This presents a challenge for their evaluation, especially when the effect on any one patient is small. In this essay, we explore the nature of the health-care system and discuss how the empirical evidence produced within it relates to the underlying processes governing patient outcomes. We argue for an evidence synthesis framework that first models the underlying phenomena common across different health-care settings and then makes inferences regarding these phenomena from data. Bayesian methods are recommended. We provide the examples of electronic prescribing and increased consultant provision at the weekend.

Scientific summary

Decisions to adopt new health technologies rely on evidence of their effectiveness along with their costs. For targeted clinical interventions, this evidence may come from randomised studies with well-defined end points. The effects of structural interventions or policy changes in health-care services are not as easily measured, as they are often disparate, affect multiple processes and end points, and may only be small in any one patient. The evaluation of structural interventions and policies therefore requires the synthesis of multiple forms of evidence from across the causal pathway that links the intervention to the outcomes that are relevant for the decision-making process.

The health-care system is a complex system that features multiple, interacting causal processes, emergent behaviours at different levels and non-linear responses to change. However, there are phenomena that are consistent across different health-care settings, and the causal processes by which an intervention may affect patient clinical outcomes are generally understood. These phenomena are distinct from the data from which they are inferred. These data may take multiple forms and be subject to many sources of bias and error. A researcher can, through literature review and expert consultation, construct a qualitative causal model of the phenomena of interest. This provides a framework both for identifying the relevant

literature and for the development of estimators of the effect of interest. Well-established evidence synthesis tools such as meta-analysis and bias modelling can then be used to make inferences about the phenomena of interest and their relationships. A Bayesian methodology is best suited to this form of research, as it permits the propagation of uncertainty through models, fits naturally in a decision-making framework and allows researchers to update results when new information becomes available.

Ongoing developments in evidence synthesis, such as methods for rapid reviews, bias modelling, and synthesising qualitative and quantitative evidence, will improve the evaluation of structural interventions. Many interventions converge on the same causal processes; research can be optimally targeted at understanding such pathways to facilitate future evaluations.

The purpose of evidence synthesis

A wide range of policies and interventions is available to the health-care system. Each of these has a potential effect on patient health and quality-of-life outcomes and a decision must be made whether or not to implement each policy or intervention. Any choice, including doing nothing, has an opportunity cost, which is the best outcome that could have been achieved using the same resources. The appropriate policy decision must, therefore, turn on the basis of whatever evidence is available.

The purpose of evidence synthesis is often to inform clinical decisions.[1] For example, the National Institute for Health and Care Excellence in England uses systematic reviews and meta-analyses to produce clinical guidelines,[2] often based on cost-effectiveness analyses, which may themselves include evidence syntheses.[3] Such evidence synthesis is generally only across studies. In the more complex case of health services research, synthesis takes place both within and between studies. The reason for this is to be found in the complex nature of casual pathways that exist in health services research.

A feature of health service interventions is that they propagate themselves across the health-care system. That is to say, there is a causal chain that may involve many mediating variables between an intervention and its effects on patients.[4] It follows that changes in the system which are captured by intervening variables are important in interpreting and explaining the effect of service interventions. A simple example of such a causal chain is shown in *Figure 1.1*. The system is conceptualised as having different levels which may begin at the patient level and end in the Department of Health. Within this framework, interventions could be classified according to the ‘level’ at which they act on the system. Clinical interventions, such as clinical guidelines, are designed to impact on clinicians; generic interventions, such as a human resources policy, are designed to impact at a managerial level; and policy interventions, such as the method of reimbursement, may be designed to interact across many organisations.

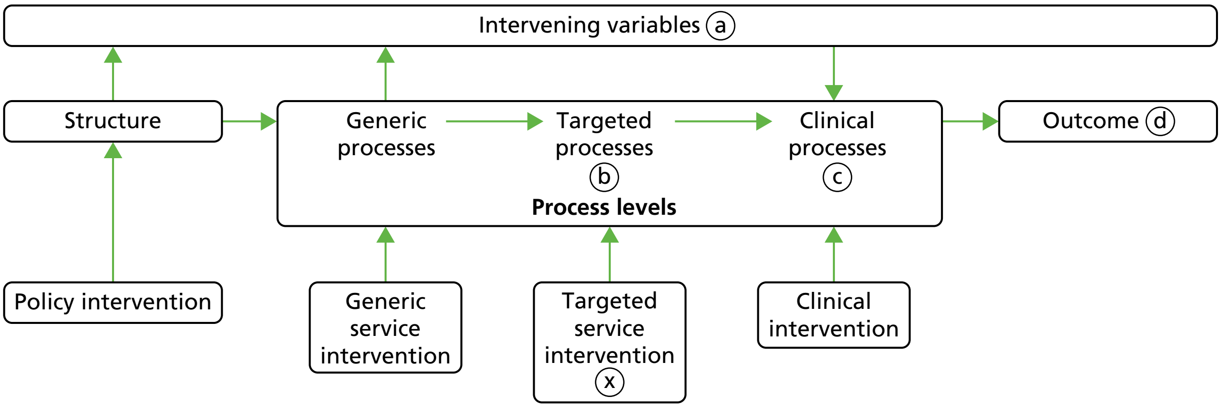

FIGURE 1.1 Causal chain showing how interventions at different levels may impact on downstream processes and outcomes. Reproduced from Evaluating policy and service interventions: framework to guide selection and interpretation of study end points, Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor CA, Barach P, vol. 341, p. c4413, 2016,[4] with permission from BMJ Publishing Group Ltd.

The purpose of an evidence synthesis is to piece together the various observations from each level that have been made of a particular system to gain knowledge of the relationships between the variables in this system. An individual study may provide insight into a number of aspects of the health-care system. It may observe a number of relevant variables that impact on the effect of an intervention across the causal chain. But, as in health technology assessment, this evidence must be assimilated in combination with other studies examining the same phenomena. There may be biases within any study, including randomised controlled trials (RCTs). Indeed, within any one study there may be biases as a result of study design, problems with implementation or interacting factors from other parts of the organisation, each of which needs to be considered when making any inferences. Between studies there may be differences in the observed relationships between variables, whether through natural variation or differing institutional contexts, and there may be publication biases. Thus, evidence synthesis needs to takes place both within and across studies.

The decision-maker needs to understand how an intervention functions and to be able to predict its effects. The process by which it functions may be relatively simple, such as how a particular medication affects the risk of mortality, or more complex, such as how a policy within a hospital affects patient length of stay, for example. In the former case, the intervention and the outcome can often be measured directly, along with any relevant variables that may causally related, in an experimental setting. In the latter case, the effect on any one patient may be too small, the duration between implementing the intervention and observing changes in patient outcomes may be too long, or there may be co-occurring changes to the institution to warrant any single study and make reliable inferences from it; in these cases the evidence is generally limited to other more upstream outcomes.[4] In this case, evidence produced from across the causal pathway would need to be synthesised to quantify the effect of interest.

What is clear is that the nature of the system producing the evidence observed needs to be understood for inferences to be made from such evidence. A 'black box' approach is not recommended, for reasons we explain later. For large systems, such as the health-care system, inference often takes place in a piecemeal fashion, with many small studies each examining parts of the whole system. There may be a number of different candidate causal models which explain the phenomena of interest that may be empirically indistinguishable from one another.[5] Specific knowledge of the health-care system is required for development of a valid causal theory. Thus, all forms of evidence produced may help to clarify the question under consideration, from ethnographic and qualitative evidence to quantitative evidence, both observational and experimental.

Decisions have been emphasised as an important motivating reason for conducting an evidence synthesis. Bayesian inference works more naturally with decision analysis. A decision-maker may want to know, once we have taken the evidence into account, what the probability is that an intervention is effective, or what the odds are that the value of an effectiveness parameter lies in a particular region. A Frequentist must remain tongue-tied in the face of such a question, whereas Bayesian methods results can be interpreted in this way. The choice of methods should therefore reflect not only what is being studied but also the reason for which it is being studied.

The nature of the system

Health service institutions are complex systems. It is important here to distinguish between a complex system and a complex intervention. The latter describes an intervention that may be composed of multiple interacting components which may differ depending on the context in which it is used; one example of this is an electronic prescribing system for hospitals.[6] Guidance already exists for the evaluation of complex interventions.[7] A complex system, on the other hand, describes a set of dynamic properties of a system. These properties make analysis and evaluation more challenging, and methods usually used to evaluate simple, clinical interventions may not be appropriate.[8] (Issues of complexity in evaluation are further discussed in *Essay 6*.)

Let us consider why a health-care institution is a complex system.[9] It comprises many interacting, casual processes. As stated above, there are multiple levels of the system, so patterns of behaviour can be observed at the individual patient and clinician level, or at more aggregate levels such as the ward or even hospital level. There are emergent processes – by which we mean that the behaviour of the system, when viewed at an aggregate level, arises from the interaction of agents at a lower level, despite those agents not exhibiting the same behaviours. For example, increases in waiting times or systematic failures may occur in accident and emergency departments despite the behaviour of all the clinical staff and their interactions with patients remaining the same. Non-linear relationships exist in health care and the output may be greater or less than the sum of its parts. Small changes to processes whereby components of the health-care system interact, such as an information technology system to identify medication errors in general practice,[10] may have large effects on patients. This may then improve patient outcomes, freeing up clinician time and other resources, leading to further improvements for other patients: these are spillover effects.

One may view the complexity of a health-care system as prohibiting successful modelling of that system, but this view would ignore the successes of other fields which study complex systems. For example, to model infectious disease epidemiology, we generally use simple models at an aggregate level, such as the Susceptible, Infected, Recovered model.[11] From a more general perspective, the methods of biology are different from those of physics despite biological objects comprising physical objects such as atoms and molecules. Biology enjoys success despite being a study of complex systems.

The difficulty of conducting experiments differs across various types of complex system. In biology it may be possible to conduct experiments at the system level: cells or individual people can be randomised to different conditions. In climate science, on the other hand, the system as a whole cannot be subject to experimentation, although experiments could be conducted on parts of the system. Health services offer an intermediate position; sometimes experiments, specifically RCTs, are possible, especially for interventions that interact 'close to the patient' (e.g. targeted service interventions in *Figure 1.1*). The repertoire of types of trial for health services research is further discussed in *Essay 2* in this volume. For example, there are many hundreds of trials of methods to improve adherence to quality standards (e.g. Flodgren *et al.*[12,13]). Experiments have even been conducted at the level of whole hospitals or wards (e.g. Hillman *et al.*,[14] Cumming *et al.*[15]). Nevertheless, there are many circumstances in which experiments are just not possible and have not been done. Alas, even when experiments have been done, it is still necessary both to understand why or how the intervention succeeded or failed, and to link the outcomes in the observed study to many other relevant patient clinical outcomes. And, to make rational and consistent scientific judgements about the effects of interventions on phenomena outside of the study design, or in a new time or place, an understanding of mechanisms is required.

The aim of models is to predict and explain phenomena in the health-care system. Each element in the model is representative of a phenomenon, which we can describe as a stable feature of the health-care system. Phenomena here are distinguished from the data from which they are inferred.[16] This distinction leads to a number of important conclusions relevant to evidence synthesis. Phenomena are generally stable and are the result of the confluence of a manageable number of causal factors, whereas data are noisy measures of the phenomena that are generated by a very large number of factors, including measurement error and bias.

Consider the model presented in *Figure 1.2*, which shows the relatively simple example of how an electronic prescribing system may impact on patient clinical outcomes. The phenomenon of an adverse drug event in this model is caused by a medication error made in the process of health-care delivery. There may be other relevant causal factors, such as the presence or absence of a monitoring system, but the underlying theory is fairly straightforward. The data from which the presence of an adverse drug event is inferred may be very noisy. Typically, adverse drug events are identified by case note review; different reviewers may have different thresholds for what is considered an adverse drug event, the quality of case notes might differ from hospital to hospital, and so forth. We can, therefore, distinguish between

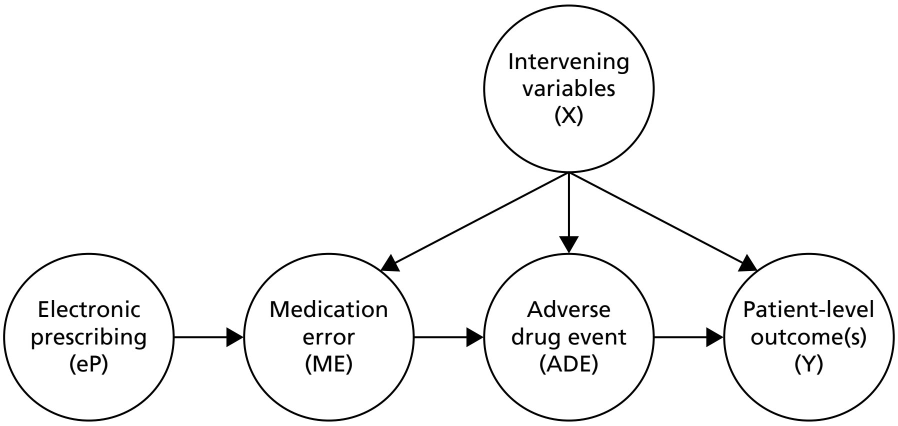
Intervening variables (X)
Electronic prescribing (eP)
Medication error (ME)
Adverse drug event (ADE)
Patient-level outcome(s) (Y)

FIGURE 1.2 Qualitative causal model showing the effects of an electronic prescribing system on patient-level outcomes.

assessing the reliability of data and explaining the underlying phenomenon.[16] This may be considered a viewpoint from a realist philosophy of science. For example, in the evaluation of a patient safety initiative, Benning *et al.*[17,18] measured errors and adverse events in multiple institutions. Multiple expert reviewers were used in the case note review, and there were controls for seasonal effects as well as reviewer learning and fatigue. However, these very real issues of data quality should not be conflicted with the phenomena to which they relate: in this case the link between intervention components and intervening variables and the link between intervening variables, clinical processes and adverse events (see *Figure 1.1*). We return to the question of assessing data reliability later and of methods for dealing with biases. To quote Bogen and Woodward:[16]

> *In undertaking to explain phenomena rather than data, a scientist can avoid having to tell an enormous number of independent, highly local, and idiosyncratic causal stories involving the (often inaccessible and intractable) details of specific experimental and observational contexts. He can focus instead on what is constant and stable across different contexts. This opens up the possibility of explaining a wide range of cases in terms of a few factors or general principles. It also facilitates derivability and the systematic exhibition of dependency-relations.*

The models can and should reflect important aspects of the system, such as why there are non-linearities in the system, but ultimately we are trying to explain why an intervention works and predict its effects. The function of the data is to help with this task.

There are both observable and unobservable processes governing a health-care system. For example, the mortality rate may be easily measurable, but levels of staff morale are not. The question to the researcher then is how to gain knowledge of these processes and their structure in order to develop a model. We take the point of view that abduction is how knowledge is gained: the theory that is inferred from observation is that which is most likely, and most simple.[19]

To illustrate the abductive process, Lipton provides the example of Ignaz Semmelweis.[19] Semmelweis wanted to find the cause of childbed (puerperal) fever in order to prevent the high rates of mortality it was causing in the Viennese hospital where he worked in the 1840s. Semmelweis had competing hypotheses about why the incidence of childbed fever was much higher in one maternity division than the other. The accepted explanation at the time was one of ‘epidemic influences’ that descended over entire districts, but that could not explain the difference between the divisions. Other explanations considered included that medical students and midwives received their training in different divisions, that a priest giving last rites had to always pass through one division to get to the other, and that women were delivered on their sides in one division but on their backs in the other. After he observed his colleague die from a disease resembling childbed fever after puncturing a finger during an autopsy, Semmelweis inferred that 'cadaveric matter' was the cause. To test this he had medical students disinfect their hands after performing autopsies and the mortality rate dropped significantly.

Semmelweis did not require knowledge of the germ theory of disease to produce knowledge of how the high mortality rate was being caused. Nor did he require knowledge of how he could measure the mortality rate or even an agent-based model of clinicians on the ward. Contrastive explanation and the weighing up the probabilities of hypotheses can help us to understand complex systems. Simple models may be objected to on the basis that they do not capture the minutiae of reality. But do we need to know how being tired can cause a clinician to make an error to know that clinicians sometimes make errors and that some of those errors cause patients harm and that an intervention that causes fewer errors reduces patient harm? Indeed, the underlying casual model may be fairly simple; it is the processes governing the data we observe that may be highly complex and context dependent.

With the use of appropriate models and methods, evidence can be synthesised to gain understanding of the health-care system, and to make predictions about the effects of policies and interventions. To reinstate, many forms of evidence are required to make accurate inference and knowledge can be gained about phenomena in a model. Moreover, this line of argument also leads us further to support a Bayesian perspective, as it permits us to weigh up the probabilities of different hypotheses and, in the last analysis, even the models themselves, allowing us to predict what happens when we intervene.[19]

Methods of evidence synthesis

Thus far, we have discussed both why evidence syntheses might take place within health services research and what the nature of the system producing the evidence may be. How does this translate into methodology? We consider three main steps: theory, data collection and evaluation, and evidence synthesis. These steps hopefully provide a useful and logical method by which a scientifically valid and useful evidence synthesis may be conducted. Where methods are lacking or underdeveloped, we have tried to offer tentative suggestions.

Theory

First, an underlying theory for the phenomenon or phenomena being studied needs to be explicated. This theory comes in the form of a model of how the system of interest functions at the level of interest. Consider again the diagram in *Figure 1.1* which provides a taxonomy of interventions depending on the level at which they interact with the system. This will act as our starting block for developing a model. For a clinical intervention the model may be very simple. *Figure 1.3* shows the example of clopidogrel (Plavix®, Bristol-Myers Squibb) for the prevention of myocardial infarction. The intervention directly acts on a single, easy-to-measure patient outcome (although other outcomes, such as costs and quality of life, may equally be of interest). Certain intervening variables may affect the chance of experiencing a myocardial infarction and some may, in practice, affect how likely it is for a patient to receive or comply with the treatment. Some of these intervening variables may be unobservable. Nevertheless, in a well-conducted RCT, these intervening variables should be the same in both control and treatment groups. The role of intervening variables cannot be overlooked, however, because they may interact with how the intervention works in a clinical context.

A generic service intervention may have a more complicated model. *Figure 1.2* shows again the relatively simple example of how an electronic prescribing system may impact on patient clinical outcomes. The aim of such a system is to reduce the medication errors that occur, and contingent patient harm in the form of adverse drug events. These in turn may have effects on patient clinical outcomes such as death or quality of life. Furthermore, the effectiveness of the intervention may depend on some other set of intervening factors, for example the ability of physicians to understand and follow clinical advice from the computer.[20] Similarly, other intervening factors may affect the risk of experiencing a medication error, adverse drug

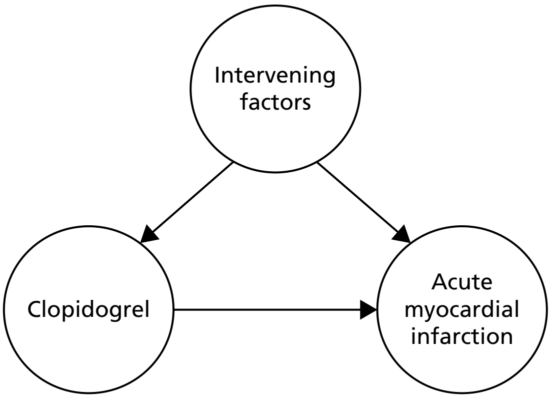

FIGURE 1.3 Qualitative causal model for the effects of clopidogrel on acute myocardial infarction.

event and clinical outcomes, such as the skill and morale of clinical staff and the availability of other interventions in the hospital environment. For many generic service and policy interventions, local hospital culture, the presence or absence of mediating factors and patient case mix are all intervening factors.[21] Indeed, such models may become quite complex as the scope of the intervention grows. *Figure 1.4* shows a candidate model for the effects of increasing consultant-to-patient ratios.

At a more practical level, how are these models interpreted? They encode our assumptions about the conditional dependencies between variables making them clear to any reader. The models presented in *Figures 1.2–1.4* are Bayesian causal networks represented as directed acyclic graphs, a common form of model used in a variety of fields, including epidemiology, statistics, philosophy and economics.[22–24] These models provide an economical representation of joint probability functions and facilitate efficient inferences from multiple observations.[23] More specifically, the model encodes the conditional dependencies between a set of random variables, some of which may be unobservable. In *Figure 1.2*, the model represents the relationships between the variables and encodes probabilistic statements about these relationships, such that we would not observe an effect of electronic prescribing on adverse drug events statistically if we condition on medication errors. On the basis of these models, we can derive, for example,

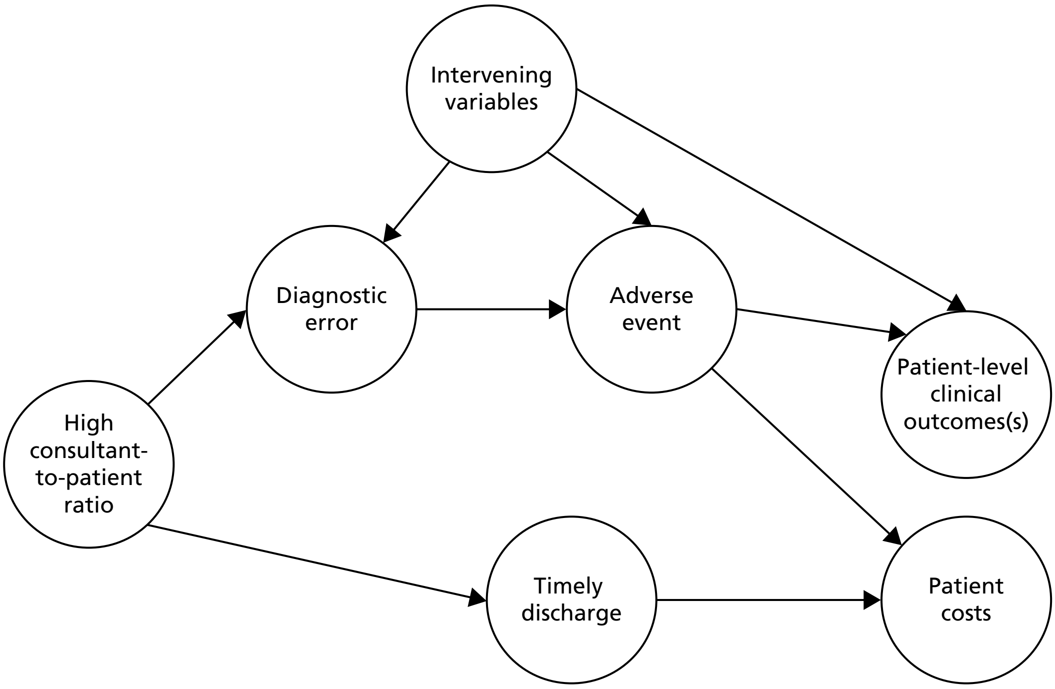

FIGURE 1.4 Qualitative causal model.

the risk of mortality for a patient treated in a hospital with an electronic prescribing system in terms of the intervening variables. This is important in order to derive estimates from multiple studies from across the causal pathway. We will return to this topic in the section *Evidence synthesis*.

It may not seem satisfactory for a researcher, in isolation, to develop a model. The researcher may lack important understanding of the causal factors involved. Agents within the system, such as doctors, nurses and managers, have important first-hand knowledge of the system. Through expert focus groups, ethnography and other qualitative means, theory can be developed and then reflected in the model. An iterative process can be used, whereby a model may be presented back to experts and clinical staff and refined further.

Developing a model iteratively raises the issue of the level of granularity and choice of variables appropriate for the model. Granularity refers to the grouping of variables. For example, should we include a general 'medication errors' variable in the model or a number of more specific variables relating to different types of medication error, such as dosing errors or allergies, or even more specific such as not prescribing low-molecular-weight heparin, on the basis of patient weight or prescribing penicillin to a patient who has already demonstrated an allergy? For both points, we suggest two criteria: (1) does the model better explain the phenomenon (as opposed to the data) or permit more precise predictions; and (2) does the model facilitate inferences that can be made about the phenomena from data? To illustrate these questions, we consider again medication errors: (1) more granularity may more precisely explain how an electronic prescribing platform reduces medication errors, as it may only act through certain types of error. However, it may hinder our ability to make predictions, particularly quantitative predictions, as the relationship between each of the medication errors and their relationship with adverse drug events would need to be explicated unless some fairly strong assumptions were made, such as independence between different types of medication errors. For (2), there may be far fewer data available for each type of medication error than for medication errors overall, and these data may be less reliable given low event rates of very specific events. This may prohibit inferences about the phenomena in a more granular model. And, as both the more and the less granular models are positing the same causal model in essence, we are not committing an error of conflating the data with the model. An example of grouping variables is given in the context of an intervention to reduce adverse events following discharge from hospital.[25]

It may be questioned why we are concerned specifically with modelling all the way to patient outcomes rather than being satisfied with more upstream outcomes. The response to this is a general health economic concern. Given the limited resources of the health-care system and potentially unlimited health-care needs, the portfolio of policies and interventions that are invested in should maximise the returns (equity considerations aside); but we are left with the issue of comparing interventions that have a wide range of potential outcomes. Health benefits may be compared on the basis of a 'natural' unit, such as deaths averted. However, this may be too narrow and capture only a small range of possible outcomes, which is especially likely to be true in the case of policy or generic service interventions. In this case, the quality-adjusted life-year is often used, which accounts for changes to both quality and length of life. For these reasons, it is important to determine the effect of these interventions on patient outcomes. Once a satisfactory model for doing so has been developed, the next step is to identify and evaluate the available evidence.

Identifying and evaluating evidence in the literature

Here, we will only provide an overview of methods that may be of practical use for the evidence synthesis methods we describe.

The question for the literature search is what needs to be found. The model developed for the synthesis guides the search for literature: data from which the phenomena and conditional probabilities in the model can be inferred are required. For the model in *Figure 1.3*, the only essential studies are those which have examined the change in risk of acute myocardial infarction conditional on clopidogrel therapy.[26]

For the model in *Figure 1.2*, a greater number of searches are required: the risk of experiencing a medication error conditional on there being an electronic prescribing platform, the risk of experiencing an adverse drug event conditional on electronic prescribing, the risk of various patient clinical outcomes conditional on electronic prescribing, the relationship between medication errors and adverse drug events, and the relationship between adverse drug events and patient clinical outcomes and costs. The first three of these searches concern the relationship between electronic prescribing and the various variables that may be considered outcomes. As previously discussed, the more downstream an outcome is from the intervention, the smaller the potential effect and hence the greater the required sample size required to detect such an effect in a study. This is likely to translate into fewer studies the greater the distance there is between the intervention and outcomes on the causal chain, and the greater the noise associated with the result. For example, a recent systematic review and meta-analysis of computerised physician order entry (CPOE) systems, which are an important component of electronic prescribing platforms, found 16 studies taking medication errors as their end point, and six that addressed adverse drug events. No studies were identified that looked directly at patient clinical outcomes.[27] Furthermore, these were just the quantitative studies; qualitative studies were not considered. Methods for identifying relevant qualitative studies differ somewhat from those used to identify quantitative studies, but are well established.[28,29]

As the model becomes more complex, such as that shown in *Figure 1.4*, more searches are required. This may present a large burden in terms of time and resources for a researcher. In many cases, interventions at the generic service intervention level or policy level converge on the same processes, namely reducing patient harm and costs through reducing adverse events. As a result, the results from a literature search of this relationship can be used in many syntheses; indeed, we are conducting such a review currently.[30]

Advances are being made in improving the efficiency with which literature reviews are conducted. For example, Tsafnat *et al.*[31] describe automated systematic review procedures and examine the impact such an automated process may have on each of the stages of the systematic review. However, much of this technology focuses on the review of RCTs, which are generally reported more consistently than observational studies which are likely to constitute much of the evidence for policy and generic service interventions. Indeed, many rapid-review methodologies still require the searching of multiple databases and manual review of retrieved abstracts.[32,33]

Once relevant studies have been identified, the next step is to assess them for quality. Quality assessment criteria have been widely established for both RCTs and observational studies. For observational studies there is the Newcastle–Ottawa scale,[34] and for RCTs there is the Cochrane Risk of Bias tool.[1] This tool is designed to facilitate the researcher to identify where there is a high risk of bias in various aspects of the trial design and conduct, such as the randomisation and allocation. However, the question then remains concerning what we should do with studies that may, potentially, be biased; we return to this question shortly. The key point is that the studies need to be assessed for their reliability for making inferences about the underlying phenomenon.

There is widely considered to be a hierarchy of evidence, with certain study designs providing more reliable evidence of effectiveness than others. Setting aside systematic reviews and meta-analyses, the RCT is generally considered to be the top of the hierarchy as the experimental design allows for the control of both observed and unobserved confounding factors and should generally be easily replicable. In an ideal world, all interventions about which there is uncertainty surrounding the existence or magnitude of an effect would be assessed in such an experimental setting, but this is obviously not possible for both practical and ethical reasons. There are many who eschew non-experimental evidence. Observational evidence is often argued to be 'too biased' to make important health policy decisions. But this would be to take an extreme position. Under the right conditions, a well-conducted observational study can produce unbiased results. Indeed, these studies may be more reliable than a poorly conducted RCT. Differential dropout, for example, can lead to a large bias.[35] A recent Cochrane review comparing treatment effects reported in observational studies with RCTs found that, 'on average, there is little evidence for significant effect estimate differences between observational studies and RCTs, regardless of specific observational study design, heterogeneity, or inclusion of studies of pharmacological interventions'.[36] That said, in some cases there are considerable differences between the results of randomised and non-randomised studies of the same intervention.[37,38]

Bias is an important concern for all study types. For non-randomised studies, case-mix adjustment is demonstrably imperfect and in certain cases can even worsen potential biases.[39] Discarding any non-randomised studies, though, is extreme. Accepting these studies at face value may also be imprudent. A middle ground is perhaps more satisfactory: modelling the biases in studies. Turner *et al.*[40] describe a method to model bias in evidence syntheses. They consider both internal biases, those biases which may cause the results of the study to deviate from those from a perfectly conducted study, and external biases, which may undermine generalisability to the target population of interest. The authors provide a method of adjusting study results for both internal and external additive and proportional biases. To determine the magnitude of the biases, the authors discuss a method of expert elicitation, where a number of independent reviewers provide their beliefs about the potential size of effect that might be observed as a result of bias if there was no intervention effect. A number of different categories are considered, similar to those in the Cochrane Risk of Bias tool, and include selection, performance and allocation biases. Sterne *et al.*[41] discuss methods of dealing with publication biases that may be apparent in the literature.

Deciding which studies to include is a topic of perennial interest in evidence synthesis, even after modelling potential biases. Data may need to be lumped together or split,[42] and the impact of these decisions should be explored through sensitivity analysis.[43] For example, the sensitivity of results to the inclusion of low-quality studies should be explored.[1]

Evidence synthesis

At this stage we have identified the available evidence to 'populate' our causal model. To synthesise this evidence we need to derive an estimator for the effect of interest from our model. We will first consider only quantitative evidence and then discuss the inclusion of qualitative evidence. The effect of interest is the average treatment effect of the intervention. Considering the model in *Figure 1.2*, and assuming that we are only interested in mortality as the outcome, Y, then the average treatment effect is

$$P(Y|eP = 1) - P(Y|eP = 0), \tag{1}$$

that is, the probability of mortality for a patient treated in a hospital with an electronic prescribing system minus the probability of mortality for a patient in a hospital without an electronic prescribing system. Following Pearl,[23] we can derive a non-parametric estimator of this effect.

The intervening variables, X, have been included in the model as they were deemed of scientific interest. We may be interested in deriving the conditional effect $P(Y|eP = 1, X = x)$, for example. In the above equation, we are calculating the effect of electronic prescribing averaged over the possible values of X. However, they are not strictly relevant in this model to the causal process by which electronic prescribing has its effect. It may be decided to leave out these variables as the results from the identified studies are already conditioning on these characteristics. More importantly, these data may not be available or the studies that investigate them may be underpowered.[44] This presents a problem when these same factors may lead to a different probability of adoption of an electronic prescribing system in practice. In this case it may still be possible to identify an estimator for the causal effect; Tian and Pearl determine the general conditions under which this is possible.[45]

For simplicity in this illustration we will consider the model in *Figure 1.2* but without the intervening variables, X, so that we have the simple model $eP \rightarrow ME \rightarrow ADE \rightarrow Y$. Then, using the variable symbols in *Figure 1.2*,

$$P(Y|eP = 1) - P(Y|eP = 0) = \textstyle\sum_{ADE} P(Y|ADE)P(ADE|ME = 1)[P(ME = 1|eP = 1) - P(M = 1|eP = 0)], \tag{2}$$

where we have used the fact that an adverse drug event must be preceded by a medication error (i.e. it is a preventable adverse drug event). Each of the components on the right hand side of the equation can be estimated using the studies identified. For example, the term $P(ME = 1|eP = 1) - P(M = 1|eP = 0)$ represents the absolute risk difference for a medication error with and without and electronic prescribing system. This can be estimated using standard meta-analytic techniques to combine the results from studies where this is estimated. As an illustration of this, consider the effect of CPOE on medication errors, as discussed above. Sixteen studies examine this, which are provided in Nuckols *et al.*[27] When the absolute risk differences are meta-analysed using a random effects meta-analysis,[46] they produce the results shown in *Figure 1.5* (the posterior distribution of the risk difference). To incorporate the studies which take adverse drug events as their end point, we can use the fact that

$$P(ADE|ME = 1)[P(ME = 1|eP = 1) - P(M = 1|eP = 0)] = P(ADE|eP = 1) - P(ADE|eP = 0). \qquad (3)$$

Within-study correlations may need to be considered when both end points are measured in the same study.

A Bayesian approach has been recommended throughout this essay. Estimation of a parameter in a statistical model in Bayesian analysis involves updating a prior distribution for the parameter, which represents what is already known about the parameter, with data, which enters via a likelihood function. A prior distribution can be informative or non-informative. A non-informative prior means that no prior information is provided and may, for example, allow for all values of the parameter to be of equal probability. A posterior distribution for the parameter can then be determined from the prior and likelihood. This posterior distribution represents our subjective uncertainty about the value of the parameter. Methods for the synthesis of both qualitative and quantitative evidence using Bayesian methods have been previously discussed in the context of health services research or policy, although

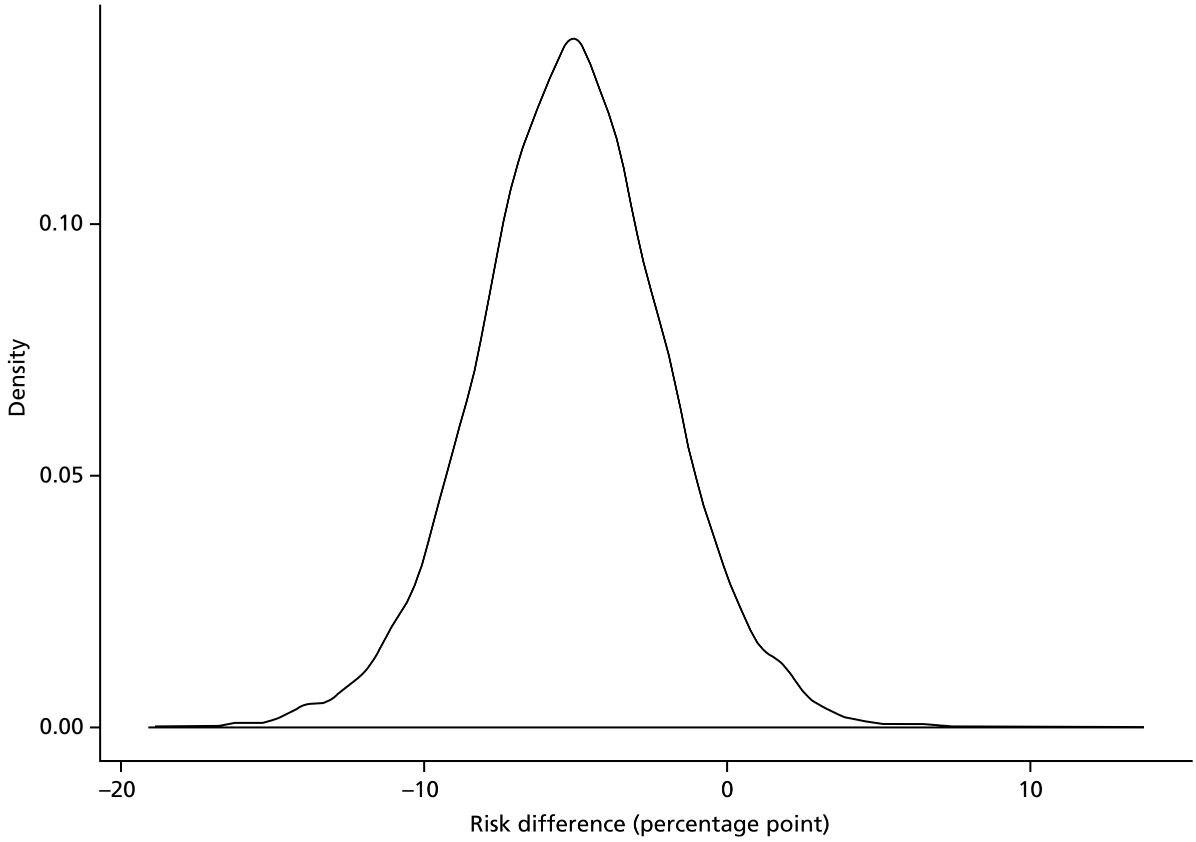

FIGURE 1.5 Posterior distribution from a random-effects meta-analysis of the absolute effect of CPOE on the risk of a patient experiencing a medication error compared with prescribing with a paper-based system.

methods have not been well developed. Roberts *et al.*[47] use qualitative evidence to generate an informative prior using expert elicitation methods. This is then updated with quantitative evidence. Voils *et al.*,[48] on the other hand, generate quantitative data from qualitative studies by attempting to determine the frequency of an association from individual reports in each study and then update a non-informative prior. In both cases, the qualitative and quantitative evidence can be used to infer the values of interest, and both provide relevant information.

Expert elicitation methods are often used to generate prior distributions in Bayesian analyses.[49,50] The example of electronic prescribing also highlights another way in which they may be of use. The only studies providing relevant evidence may be of a similar but not identical intervention to the one of interest. CPOE is only one component of a full electronic prescribing system, yet only studies examining CPOE are available. A prior distribution for the parameters in the electronic prescribing model may be elicited from experts who are asked to extrapolate from the evidence of CPOE.

Ultimately, in whichever way the various forms of evidence are combined, the effect of interest can then be evaluated over the posterior distributions of each of the parameters. Or, if no data are available, it may be evaluated over the prior distributions even if some of these are non-informative.

Decision analysis

It was emphasised at the beginning of the essay that policy decisions are often the purpose of evidence syntheses. Bayesian decision analysis provides methods to determine the optimal decision within a normative utilitarian framework.[51,52] A loss function is specified based on a utility function which represents the losses to a decision-maker. The benefits of an intervention can be determined using the methods described in this essay: the model and evidence synthesis can take place within a decision model.

Conclusions

In this essay we have provided a framework for evidence synthesis in health services research involving synthesis of external evidence from the literature and evidence internal to a study relating to salient phenomena contributing to the link between cause and effect. The evaluation of generic service interventions and policies within health systems is often hampered by the fact that the effect of these interventions on any one patient is often very small. Sample sizes required to detect such an effect may be prohibitively large, and, unless the study is perfectly conducted, the magnitude of the bias may overwhelm the size of the effect. It also may not be feasible, ethical or even necessary to conduct a randomised trial. Synthesising evidence from across the causal pathway may provide a method of estimating the effects of these interventions on the patient outcomes of interest.

Many forms of evidence are available from which the effects of interest can be inferred. However, in many cases there may be a risk of bias. This bias may arise anywhere from the conceptual stage in choosing which interventions to study right through to the publication of results where negative findings may not be published. We have discussed methods to model such biases. The Bayesian framework we have described also permits the synthesis of both qualitative and quantitative evidence: the most likely values for the phenomena described by the causal model can be 'triangulated' from the available evidence.

The methods described here combine many different methods elaborated elsewhere. However, in some cases further research is required to establish the optimal techniques. For example, there is no consensus on the best method for the integration of both qualitative and quantitative methods. Similarly, further research is required to elucidate the causal pathways present in health-care institutions. This essay is intended to provide a foundation onto which these techniques can be built, enabling estimation of the effects of generic service and policy interventions.

Acknowledgements

The authors would like to thank Professor Terri Piggott (Professor in Research Methodology), Professor Mark Petticrew (Professor in Public Health Evaluation) and Professor David Jones (Professor in Medical Statistics) for their input to this essay as well as discussants at the Evaluate Conference 2015.

Contributions of authors

Samuel I Watson (Research Fellow in Health Economics) and **Richard J Lilford** (Professor, Public Health) developed the essay and examples contained herein.

Samuel I Watson prepared the first draft and both authors contributed to each subsequent draft.

Both authors approved the final version of the manuscript.

References

1. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928. http://dx.doi.org/10.1136/bmj.d5928
2. National Institute for Health and Care Excellence. *The Guidelines Manual*. 2009. URL: www.nice.org.uk/article/pmg6/ (accessed 23 February 2016).
3. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med* 2003;**22**:3687–709. http://dx.doi.org/10.1002/sim.1586
4. Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor CA, Barach P. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *BMJ* 2010;**341**:c4413. http://dx.doi.org/10.1136/bmj.c4413
5. Mayo-Wilson C. The limits of piecemeal causal inference. *Br J Philos Sci* 2013;1–37.
6. Lilford RJ, Girling AJ, Sheikh A, Coleman JJ, Chilton PJ, Burn SL, *et al.* Protocol for evaluation of the cost-effectiveness of ePrescribing systems and candidate prototype for other related health information technologies. *BMC Health Serv Res* 2014;**14**:314. http://dx.doi.org/10.1186/1472-6963-14-314
7. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. *Developing and Evaluating Complex Interventions: New Guidance*. Medical Research Council; 2006. URL: www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/ (accessed February 2016).
8. Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. *BMJ* 2008;**336**:1281–3. http://dx.doi.org/10.1136/bmj.39569.510521.AD
9. Lipsitz LA. Understanding health care as a complex system. *JAMA* 2012;**308**:243. http://dx.doi.org/10.1001/jama.2012.7551
10. Hemming K, Chilton PJ, Lilford RJ, Avery A, Sheikh A. Bayesian cohort and cross-sectional analyses of the PINCER trial: a pharmacist-led intervention to reduce medication errors in primary care. *PLOS ONE* 2012;**7**:e38306. http://dx.doi.org/10.1371/journal.pone.0038306
11. Daley D, Gani J. *Epidemic Modelling. An Introduction*. Cambridge: Cambridge University Press; 2001.

12. Flodgren G, Pomey M-P, Taber SA, Eccles MP. Effectiveness of external inspection of compliance with standards in improving healthcare organisation behaviour, healthcare professional behaviour or patient outcomes. *Cochrane Database Syst Rev* 2011;**11**:CD008992.

13. Flodgren G, Conterno LO, Mayhew A, Omar O, Pereira CR, Shepperd S. Interventions to improve professional adherence to guidelines for prevention of device-related infections. *Cochrane Database Syst Rev* 2013;**3**:CD006559. http://dx.doi.org/10.1002/14651858.cd006559.pub2

14. Hillman K, Chen J, Cretikos M, Bellomo R, Brown D, Doig G, *et al.* Introduction of the medical emergency team (MET) system: a cluster-randomised controlled trial. *Lancet* 2005;**365**:2091–7. http://dx.doi.org/10.1016/S0140-6736(05)66733-5

15. Cumming RG, Sherrington C, Lord SR, Simpson JM, Vogler C, Cameron ID, *et al.* Cluster randomised trial of a targeted multifactorial intervention to prevent falls among older people in hospital. *BMJ* 2008;**336**:758–60. http://dx.doi.org/10.1136/bmj.39499.546030.BE

16. Bogen J, Woodward J. Saving the phenomena. *Philos Rev* 1988;**97**:303–52. http://dx.doi.org/10.2307/2185445

17. Benning A, Dixon-Woods M, Nwulu U, Ghaleb M, Dawson J, Barber N, *et al.* Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ* 2011;**342**:d199. http://dx.doi.org/10.1136/bmj.d199

18. Benning A, Ghaleb M, Suokas A, Dixon-Woods M, Dawson J, Barber N, *et al.* Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. *BMJ* 2011;**342**:d195. http://dx.doi.org/10.1136/bmj.d195

19. Lipton P. *Inference to the Best Explanation*. 2nd edn. Abingdon: Routledge; 2004.

20. Coleman JJ, Hemming K, Nightingale PG, Clark IR, Dixon-Woods M, Ferner RE, *et al.* Can an electronic prescribing system detect doctors who are more likely to make a serious prescribing error? *JRSM* 2011;**104**:208–18. http://dx.doi.org/10.1258/jrsm.2011.110061

21. Wachter RM, Pronovost P, Shekelle P. Strategies to improve patient safety: the evidence base matures. *Ann Intern Med* 2013;**158**:350–2. http://dx.doi.org/10.7326/0003-4819-158-5-201303050-00010

22. Foraita R, Spallek J, Zeeb H. *Directed Acyclic Graphs*. In Ahrens W, Pigeot I, editors. *Handbook of Epidemiology*. New York, NY: Springer; 2014. pp. 1481–517. http://dx.doi.org/10.1007/978-0-387-09834-0_65

23. Pearl J. Causality. 2nd edn. Cambridge: Cambridge University Press; 2009. http://dx.doi.org/10.1017/CBO9780511803161

24. Spiegelhalter DJ. Bayesian graphical modelling: a case-study in monitoring health outcomes. *J Roy Stat Soc C App* 2002;**47**:115–33. http://dx.doi.org/10.1111/1467-9876.00101

25. Yao GL, Novielli N, Manaseki-Holland S, Chen Y-F, van der Klink M, Barach P, *et al.* Evaluation of a predevelopment service delivery intervention: an application to improve clinical handovers. *BMJ* Qual Saf 2012;**21**(Suppl. 1):i29–38. http://dx.doi.org/10.1136/bmjqs-2012-001210

26. Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *BMJ* 2002;**324**:71–86. http://dx.doi.org/10.1136/bmj.324.7329.71

27. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, *et al.* The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. *Syst Rev* 2014;**3**:56. http://dx.doi.org/10.1186/2046-4053-3-56

28. Shaw RL, Booth A, Sutton AJ, Miller T, Smith JA, Young B, *et al.* Finding qualitative research: an evaluation of search strategies. *BMC Med Res Methodol* 2004;**4**:5. http://dx.doi.org/10.1186/1471-2288-4-5

29. Flemming K, Briggs M. Electronic searching to locate qualitative research: evaluation of three strategies. *J Adv Nurs* 2007;**57**:95–100. http://dx.doi.org/10.1111/j.1365-2648.2006.04083.x

30. Watson S, Taylor C, Chen Y. *A Systematic Review and Meta-Analysis to Identify the Health and Economic Consequences of Adverse Events at the Patient-Level*. York: University of York; 2015. URL: www.crd.york.ac.uk/PROSPERO/display_record.asp?ID = CRD42015019578 (accessed February 2016).

31. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. *Syst Rev* 2014;**3**:74. http://dx.doi.org/10.1186/2046-4053-3-74

32. Polisena J, Garritty C, Kamel C, Stevens A, Abou-Setta AM. Rapid review programs to support health care and policy decision making: a descriptive analysis of processes and methods. *Syst Rev* 2015;**4**:26. http://dx.doi.org/10.1186/s13643-015-0022-6

33. Hayden JA, Killian L, Zygmunt A, Babineau J, Martin-Misener R, Jensen JL, *et al.* Methods of a multi-faceted rapid knowledge synthesis project to inform the implementation of a new health service model: Collaborative Emergency Centres. *Syst Rev* 2015;**4**:7. http://dx.doi.org/10.1186/2046-4053-4-7

34. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M TP. *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta-Analyses.* Ottawa, ON: The Ottawa Hospital Research Institute; 2012. URL: www.ohri.ca/programs/clinical_epidemiology/oxford.asp (accessed February 2016).

35. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ* 2013;**346**:e8668. http://dx.doi.org/10.1136/bmj.e8668

36. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014;**4**:MR000034. http://dx.doi.org/10.1002/14651858.mr000034.pub2

37. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878–86. http://dx.doi.org/10.1056/NEJM200006223422506

38. Ioannidis JPA. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**:821. http://dx.doi.org/10.1001/jama.286.7.821

39. Nicholl J. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *J Epidemiol Community Heal* 2007;**61**:1010–13. http://dx.doi.org/10.1136/jech.2007.061747

40. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. *J Roy Stat Soc A Sta* 2009;**172**:21–47. http://dx.doi.org/10.1111/j.1467-985X.2008.00547.x

41. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002. http://dx.doi.org/10.1136/bmj.d4002

42. Weir MC, Grimshaw JM, Mayhew A, Fergusson D. Decisions about lumping vs. splitting of the scope of systematic reviews of complex interventions are not well justified: a case study in systematic reviews of health care professional reminders. *J Clin Epidemiol* 2012;**65**:756–63. http://dx.doi.org/10.1016/j.jclinepi.2011.12.012

43. Geneletti S, Mason A, Best N. Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only 'solution'. *Epidemiology* 2011;**22**:36–9. http://dx.doi.org/10.1097/EDE.0b013e3182003276

44. Smith PG, Day NE. The design of case–control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;**13**:356–65. http://dx.doi.org/10.1093/ije/13.3.356

45. Tian J, Pearl J. A General Identification Condition for Causal Effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press/The MIT Press; 2002. pp. 567–73.

46. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88. http://dx.doi.org/10.1016/0197-2456(86)90046-2

47. Roberts KA, Dixon-Woods M, Fitzpatrick R, Abrams KR, Jones DR. Factors affecting uptake of childhood immunisation: a Bayesian synthesis of qualitative and quantitative evidence. *Lancet* 2002;**360**:1596–9. http://dx.doi.org/10.1016/S0140-6736(02)11560-1

48. Voils C, Hassselblad V, Crandell J, Chang Y, Lee E, Sandelowski M. A Bayesian method for the synthesis of evidence from qualitative and quantitative reports: the example of antiretroviral medication adherence. *J Health Serv Res Policy* 2009;**14**:226–33. http://dx.doi.org/10.1258/jhsrp.2009.008186

49. O'Hagan A. Eliciting expert beliefs in substantial practical applications. *J Roy Stat Soc D Sta* 1998;**47**:21–35. http://dx.doi.org/10.1111/1467-9884.00114

50. Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ* 1995;**311**:1621–5. http://dx.doi.org/10.1136/bmj.311.7020.1621

51. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. 3rd edn. New York, NY: Springer; 1993.

52. Press SJ. *Subjective and Objective Bayesian Statistics*. 2nd edn. Hoboken, NJ: John Wiley and Sons; 2002. http://dx.doi.org/10.1002/9780470317105

Essay 2 Randomised controlled trials of complex interventions and large-scale transformation of services

Helen Barratt,[1] Marion Campbell,[2] Laurence Moore,[3]
Merrick Zwarenstein[4] and Peter Bower[5]

[1]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames, Department of Applied Health Research, University College London, London, UK
[2]Health Services Research Unit, University of Aberdeen, Aberdeen, UK
[3]Medical Research Council (MRC)/Chief Scientist Office (CSO) Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK
[4]Centre for Studies in Family Medicine, Department of Family Medicine, Western University, London, ON, Canada
[5]National Institute for Health Research (NIHR) School for Primary Care Research, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

Declared competing interests of authors: Helen Barratt is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care North Thames at Bart's Health NHS Trust. Laurence Moore reports that he is Scientific Advisor to and former Director of Decipher Impact Ltd. The Health Services Research Unit, University of Aberdeen, receives core funding from the Chief Scientist Office (SCO) of the Scottish Government's Health and Social Care Directorates (CSO). The Medical Research Council (MRC)/CSO Social and Public Health Sciences Unit, University of Glasgow, receives core funding from the CSO (SPHSU14) and the MRC (MC_UU_12017/14). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Published May 2016
DOI: 10.3310/hsdr04160-19

This essay should be referenced as follows:

Barratt H, Campbell M, Moore L, Zwarenstein M, Bower P. Randomised controlled trials of complex interventions and large-scale transformation of services. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 19–36.

List of figures

List of abbreviations

CONSORT	Consolidated Standards of Reporting Trials
CPRD	Clinical Practice Research Datalink
cRCT	cluster randomised controlled trial
ICC	intracluster correlation coefficient
MRC	Medical Research Council
PRECIS-2	PRagmatic-Explanatory Continuum Indicator Summary 2
RCT	randomised controlled trial

Abstract

Complex interventions and large-scale transformations of services are necessary to meet the health-care challenges of the 21st century. However, the evaluation of these types of interventions is challenging and requires methodological development.

Innovations such as cluster randomised controlled trials, stepped-wedge designs and non-randomised evaluations provide options to meet the needs of decision-makers. Adoption of theory and logic models can help clarify causal assumptions, and process evaluation can assist in understanding delivery in context. Issues of implementation must also be considered throughout intervention design and evaluation to ensure that results can be scaled for population benefit. Relevance requires evaluations conducted under real-world conditions, which in turn requires a pragmatic attitude to design. The increasing complexity of interventions and evaluations threatens the ability of researchers to meet the needs of decision-makers for rapid results. Improvements in efficiency are thus crucial, with electronic health records offering significant potential.

Scientific summary

Complex interventions and large-scale transformations of services are necessary to meet the health and social care challenges of the 21st century. The evaluation of complex interventions and large-scale transformations of services is challenging and continues to demand methodological innovation

Innovations in design (cluster randomised controlled trials, stepped-wedge designs and non-randomised evaluations) provide decision-makers with a variety of options to meet their evaluation needs. The complexity of modern interventions requires effective intervention development. The use of theory and logic models can make explicit how the intervention is expected to impact on process and outcome. The use of detailed process evaluation to understand delivery in context is also important. Relevance requires that evaluations are conducted under real-world conditions, which in turn requires changes in the design of randomised trials (e.g. greater use of pragmatic trials). Intervention design and process evaluation need to explicitly consider issues of implementation to ensure that trial results can be scaled to drive population benefit.

The increasing complexity of interventions and their evaluation threatens the speed and efficiency of the evaluation, and the ability to meet the needs of decision-makers for rapid results. Innovations such as the more effective use of electronic health records have significant potential to facilitate rapid trial delivery.

Introduction

Attempts to improve health outcomes or tackle public health problems increasingly make use of *complex interventions*. These are commonly defined as interventions that contain several interacting components.[1] Additional dimensions of complexity include the number of different behaviours required by those delivering or receiving the intervention; the range of stakeholders or structures targeted (patients, professionals, systems of care); and the degree of variation when the intervention is delivered in different contexts.

Complex interventions present a number of specific problems for evaluators. Many of these relate to the difficulty of standardising such interventions; their sensitivity to features of local context; and the length and complexity of the causal chains linking the intervention with outcome.[1] Complexity exists not only in the interactions of the different components of the intervention[2] but also in its interaction with the wider system in which the intervention is embedded.

Although a range of experimental and non-experimental approaches to evaluation have been proposed, randomised controlled trials (RCTs) remain the optimal method for obtaining unbiased estimates of effectiveness.[3] The UK Medical Research Council (MRC) has set out guidance on developing and evaluating complex interventions in trials, originally published in 2000, and revised and extended in 2008.[1] Others have further extended our understanding of relevant research methods in this area.[4]

As the pace and scope of change in health services has increased, there is also a growing need to evaluate not just complex interventions, but also *large-scale transformations of services*. These have been defined as 'interventions aimed at coordinated, system wide change affecting multiple organisations and care providers, with the goal of significant improvements in the efficiency of health care delivery, the quality of patient care, and population-level patient outcomes'.[5] 'Integrated care' for long-term conditions represents an excellent exemplar of such a transformation, involving complex packages of care as well as more profound changes in workforce configuration, service organisation and funding. The rigorous evaluation of such change represents a new level of challenge. The issues of scale and complexity in the evaluation of services are further discussed in *Essay 6* in this volume.

Evaluators also face additional pressures. Designing, delivering and understanding the operation of a complex intervention may require complex, multimethod research designs to both provide rigorous assessment and assist in the interpretation of results. Classic RCTs are traditionally long, complicated and expensive exercises and are, therefore, under significant scrutiny in an era of financial restrictions in research funding and major concerns about research waste.[6] New, more pragmatic RCT designs are receiving increasing attention.

There is pressure from policy-makers and service managers for rapid evaluation to support short- and medium-term decision-making in areas where they face pressing problems of demand and supply. While methodological rigour remains critical, increasing importance is placed on the speed with which research findings are delivered and translated into practice. Developing and testing innovations to enable implementation at 'pace and scale' is considered crucial to ensure reach and long-lasting and replicable effects. This requires greater attention to the balance between internal and external validity, building consideration of later implementation into the design of interventions at an early stage, and may set limits around the cost, time scales and sophistication of research designs.

In this essay, we first describe recent methodological developments in relation to RCTs for the evaluation of complex interventions and large-scale transformation. We review current knowledge about available approaches, comparing and contrasting methods for different scenarios. We then go on to outline issues and practical challenges that could usefully be addressed by those who conduct, fund and regulate such research.

Randomised methods

In a conventional RCT, individuals are randomly allocated between alternatives. However, complex interventions and large-scale transformations often target levels other than the individual participant. For example, whole communities may be targeted by public health programmes; other interventions may be aimed at groups of health professionals, or organisations such as general practices.

Contamination of the control group, leading to biased estimates of effect size, is a drawback of RCTs of population-level interventions.[7] This is likely to occur in situations in which different public health promotion interventions are tested within the same community, or the same clinician delivers different types of health care to patients.[8]

Cluster randomised controlled trials

Cluster randomised controlled trials (cRCTs) have emerged partly as a means of addressing such problems. The units randomised are pre-existing, natural or self-selected clusters. Cluster members all have an identifiable feature in common (such as patients in a single general practice), and outcomes are measured in all or in a representative sample. The clusters can vary widely depending on the setting, and range in size from families to entire communities.

Methodological discussion of the cRCT appears to have begun in the field of education in the 1940s,[8] but the design saw only sparse use before the 1980s. However, the last half-century has seen a steady increase in cRCTs published in the medical literature: from one per year in the 1960s to over 120 in 2008.[8] The main implication of the cRCT design is that the outcomes of individuals within the same cluster tend to be correlated, which must be accounted for in the analysis. The statistical measure of correlation is known as the intracluster correlation coefficient (ICC), defined as the proportion of variation in the outcome that can be explained by the variation between clusters. Sample size estimates need to account for this, taking into account both the size of the clusters and the ICC. Reliable estimates of ICCs are essential for robust sample size calculations. There is increased understanding of the factors that affect clustering and therefore the likely magnitude of an ICC. For example, Campbell *et al.*[9] demonstrated that ICCs are sensitive to a number of trial-related factors, particularly the setting and the type of outcome. Analysing a range of data sets, mainly from the UK, they found that ICCs were significantly higher for process than outcome variables and for outcomes in specialist settings compared with primary care. The effects of disease prevalence and trial size were less clear-cut.[9] Other studies, meanwhile, suggest that ICC is associated with prevalence for binary outcome measures.[10]

There is also growing understanding of how to calculate sample sizes for cRCTs, including scenarios in which the size of individual clusters varies.[11] This has been incorporated into standard statistical packages, and relevant online tools are also available.[12]

The third development has been the publication of reporting guidelines for cRCTs. These guidelines highlight that it is important to indicate the level at which the interventions were targeted, hypotheses were generated, randomisation was done and outcomes were measured.[13] In 2012, specific guidance was published, based on the 2010 version of the Consolidated Standards of Reporting Trials (CONSORT) statement and the 2008 CONSORT statement for the reporting of abstracts.[14] The guidance includes a checklist of items that authors should address when reporting cRCTs, in conjunction with the main CONSORT guidelines.[15]

There has also been the development of an appropriate ethical framework, *The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomised Trials*. This sets out 15 recommendations, providing guidance on the justification for a cRCT; the need for ethical review; the identification of research participants; obtaining informed consent; the assessment of benefits and harms; and the protection of vulnerable participants.[16]

Stepped-wedge trials

The stepped-wedge trial has emerged as an alternative to the standard parallel-group cRCT and is increasing in popularity for evaluating complex interventions and large-scale transformations.[17] The design includes an initial period during which no clusters are exposed to the intervention and therefore act as controls. Subsequently, at regular intervals (the ‘steps’), one cluster (or a group of clusters) is randomised to cross over from control to intervention. This process continues until all clusters have crossed over and are exposed to the intervention (*Figure 2.1*). The appeal of the design is that, at the end of the study, there will be a period when all clusters are exposed, but the order of intervention delivery is determined at random. Data collection continues throughout the study, so that each cluster contributes observations during both control and intervention observation periods.[19]

The rationale for using this approach, as well as specific methodological considerations, has been described in more detail elsewhere.[19,20] However, the principal advantage is that it may more adequately reconcile the differing needs and priorities of policy-makers and service managers on the one hand, and evaluators on the other. It has intuitive appeal, as each cluster eventually receives the intervention and acts as its own control. As a consequence, it is particularly suited to interventions in which there is a lack of evidence of effectiveness but a strong belief that they will do more good than harm,[21] such as efforts to improve hand hygiene. It also promotes the phased implementation of novel interventions, permitting time for staff training, if this is required. In addition, the researcher may study the effect of time on an intervention, such as whether or not adoption or learning is maintained.

Nevertheless, because data collection usually takes place at each time point in the study, the data burden is potentially significant. It may also take time before every cluster receives the intervention. In addition, the analysis is more complex than that required for a conventional cRCT. Sample size calculations and analysis must make allowance for both the clustered design and the confounding effect of time.[20] In a stepped-wedge design, more clusters are exposed to the intervention towards the end of the study than in its early stages. Consequently, the effect of the intervention might be confounded by any underlying temporal trends if, for example, there has been a general move towards improved patient outcomes, external to the study.[19] Of course, judging whether or not an intervention is likely to do more good than harm is complex, especially when considering wider issues such as opportunity costs. A stepped wedge may mean that all clusters are exposed to an ineffective intervention if initial estimates of effect prove unfounded.

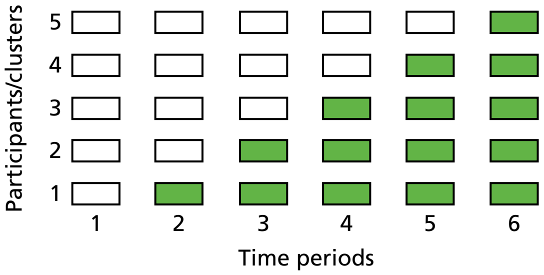

FIGURE 2.1 Stepped-wedge trial schema. Shaded cells represent intervention periods, blank cells represent control periods and each cell represents a data collection point. Reproduced from Brown and Lilford under the terms of the Creative Commons Attribution Licence (CC-BY 2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.[18]

Selecting a randomised approach

Stepped-wedge trials and cRCTs both offer a means of evaluating complex interventions and large-scale transformations. In what situation should one be preferred? Consider a hypothetical study, involving a group-based weight loss programme with internet support. Early pilot investigations suggest that the intervention is effective, with a primary outcome of weight loss after 6 months. There is interest locally in implementing the programme; eight obesity clinics have provisionally agreed to participate.

In this example, we assume a minimum duration of follow-up of 6 months, and monthly introduction of the intervention into new clusters. In the event that the intervention is effective, the final results would be available in 14 months using a stepped-wedge design. By that stage, all the study clusters would have at least 6 months' experience of using the intervention, and have been reaping the benefits. In contrast, if a parallel-arm cRCT were used, this result would have been known within 6 months. Four of the clusters would have been delivering the programme since the start of the trial. However, it would now be necessary to 'fast track' the intervention in the four control clusters.

On the other hand, consider if the intervention were not effective. With a stepped-wedge design, the final results would again be available in 14 months. However, by that point, all of the clusters would have been exposed to an intervention that does not work, and all would have been delivering it to patients for at least 6 months. Consequently, ethical scrutiny of this type of trial is important. In addition, because data collection usually takes place sequentially, the researchers would have collected a full set of trial data on 14 separate occasions (although a smaller number of data collection periods could be considered). If a cRCT were used, the results would be known within 6 months. Only half of the clusters would have been exposed to an intervention that did not work, and trial data would have been collected only once.

Key questions to be addressed before commencing such a trial include:

- What is the strength of the evidence that the intervention is more likely to do more good than harm? Considerations of costs and opportunity costs may be relevant.
- As well as effectiveness, what other issues are being assessed by the trial, such as learning and possible decay of the intervention effect?
- How quickly could an effective intervention be rolled out to control clusters if a cRCT were conducted?

The optimal conditions for a stepped wedge are likely to be when there is a lack of evidence of effectiveness but a strong belief that the intervention will do more good than harm.[20] It may also be relevant when measurement of the effect of the intervention over time is important, including evaluation of staff learning and the effect of the service 'bedding down'. In contrast, a cRCT may be more appropriate when the evidence of effectiveness is not clear-cut and rapid implementation in control clusters is possible. cRCTs have the added advantage that half of the study sample is 'protected', should the intervention not work, and the data collection burden is minimised.

In many situations, the concerns of those commissioning the evaluation may outweigh methodological issues. The constraints under which policy-makers and service managers operate may mean that a stepped wedge is the only way to gain agreement to randomisation, because all clusters will receive the intervention during the research. As a consequence, stepped-wedge designs may be particularly useful when policy dictates that an initiative will be rolled out widely, and there is scope to do this in a stepped way to enhance evaluation.

There are examples of successful cRCTs of complex policy interventions, which have successfully negotiated issues around randomisation and implementation, such as a cRCT of a scheme to provide free healthy breakfasts in Welsh primary schools.[22] Of the 608 schools invited to participate, 111 agreed to be randomised and participate in data collection activities. Stepped-wedge designs have the potential to appear more attractive to stakeholders and may encourage higher levels of participation.

Pragmatic and explanatory attitudes to trial design

The PRagmatic-Explanatory Continuum Indicator Summary 2 (PRECIS-2) tool is designed to support trial design decisions consistent with the intended purpose of a planned RCT. This tool is especially useful if the intention is to design a RCT that will have results widely applicable within usual care at whatever level of the health service this may be, from primary to tertiary care settings. PRECIS-2 consists of nine domains: eligibility criteria, recruitment, setting, organisation, flexibility (delivery), flexibility (adherence), follow-up, primary outcome and primary analysis (*Figure 2.2*). These are discussed until a consensus is reached that that particular element of the trial is designed in a way which promotes the goal of the trial, be it a real-world applicability or research designed purely to understand efficacy. The process makes research teams more aware of the range of opinions and can help to ensure that design decisions are matched to the needs of those who will ultimately use the results of the trial.[23]

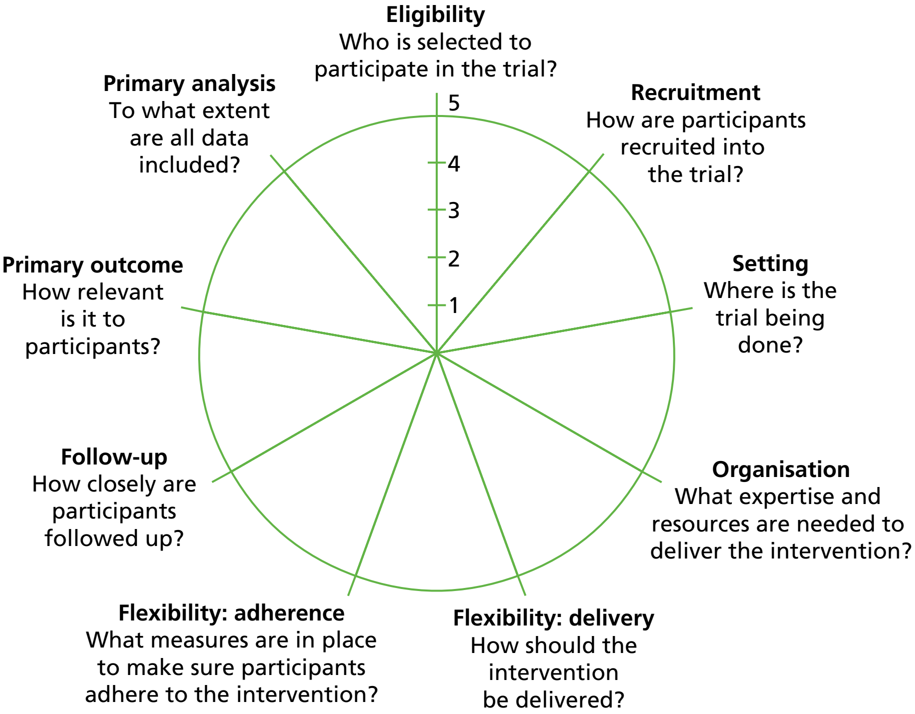
Eligibility
Who is selected to participate in the trial?
5
4
3
2
1
Recruitment
How are participants recruited into the trial?
Setting
Where is the trial being done?
Organisation
What expertise and resources are needed to deliver the intervention?
Flexibility: delivery
How should the intervention be delivered?
Flexibility: adherence
What measures are in place to make sure participants adhere to the intervention?
Follow-up
How closely are participants followed up?
Primary outcome
How relevant is it to participants?
Primary analysis
To what extent are all data included?

FIGURE 2.2 The PRECIS-2 wheel. Reproduced from The PRECIS-2 tool: designing trials that are fit for purpose, Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M, vol. 350, p. h2147, 2015,[23] with permission from BMJ Publishing Group Ltd.

Non-randomised methods

Policy-makers overseeing the rollout of a new programme may not be willing to accept any form of randomisation. Even a lack of control over the order of entry of clusters into a stepped wedge may be unacceptable. This may be for political reasons, as we have described, or may occur when the intervention is already in widespread use, or key decisions about how it will be implemented have already been taken.[1] In this scenario, the best evidence may come from a study design further down the conventional evidence hierarchy, such as a controlled before-and-after study or an interrupted time series. The revised MRC guidance cites a range of studies that have successfully evaluated complex interventions using non-randomised methods, including bans on smoking in public places,[24] the impact of advice about sleep position on the incidence of sudden infant deaths[25] and the impact of financial incentives on quality.[26]

Non-randomised methods that attempt to deal with the effects of selection, allocation and other biases include matching, conventional covariate adjustment using regression models and propensity score methods.[27] A propensity score is defined as the conditional probability of an individual receiving an intervention, given the observed covariates. It can be used to balance the covariates between study groups, and therefore reduce the risk of bias.

This approach is potentially useful for the evaluation of complex interventions where randomisation is not feasible or desirable. However, further work is required to enable better adjustment for case-mix differences. For example, confounding by indication offers challenges to the analyst. In this situation, allocation to treatment is not otherwise ignorable but instead is subject to some latent (unrecognised or unmeasured) process associated with those who are treated, for example when skilled clinicians use their expert judgement to decide whether or not to treat a patient. This judgement may include criteria describing illness severity or patient frailty not included in the propensity score or, more likely, not formally measured. Methods such as propensity scores account only for known and observed patient characteristics.[28] Methods that are better able to deal with unobserved factors include differences-in-differences analysis, instrumental variable analysis and regression discontinuity designs.[29] There is further discussion of non-randomised designs in *Essay 3* in this volume.

Of course, selection bias is only one form of bias in evaluations of complex interventions and large-scale transformations. There are other aspects of good RCT design and delivery that can mitigate other sources of bias (such as protocols, pre-specified primary outcomes and analytic plans) which can be applied to both randomised and non-randomised designs.[30]

Many non-randomised designs use routine data, which can have significant limitations in terms of accuracy and scope. Such data may lack patient-oriented outcomes such as quality of life and patient experience. A recent assessment of a primary care policy initiative in the North West to improve access to care[31] was able to use routinely collected patient experience data from the General Practice Patient Survey to provide a more rounded assessment of outcomes, alongside admissions and health-care utilisation data.

Wider practical issues

In the final section of this essay, we outline practical issues of relevance to those seeking to implement RCTs of complex interventions and large-scale transformations.

Challenges in intervention design

As the revised MRC complex interventions guidance highlights, best practice is to develop complex interventions systematically and then to test them via pilot studies, before moving on to an exploratory and then a definitive evaluation.[1] In practice, however, policy implementation often does not follow this model, leaving researchers with less opportunity to influence intervention design or delivery. An example from England would be the government's population-wide prevention programme for cardiovascular disease (the NHS Health Check programme), which was rolled out despite a number of uncertainties about the evidence base underpinning this approach and how best to encourage the necessary uptake among patients to achieve the expected gains.[32] Ideally, all intervention development should begin with identification of a relevant evidence base. This will often need to be supplemented by the application of appropriate theory, which may lead to greater probability of an effective intervention than a purely empirical or pragmatic approach.[33]

Even in cases where researchers cannot influence the design and implementation of an intervention, they can usefully help decision-makers to develop a theoretical understanding of the likely process of change by articulating the rationale behind it. This includes the changes that might be expected, and how change is to be achieved. A 'logic model' can help clarify causal assumptions, which may be drawn from behavioural, social science or organisational theory. Making explicit the causal assumptions and mechanisms of effect can allow external scrutiny of its plausibility and help evaluators decide which aspects of the intervention or its context to prioritise for investigation.[3]

Increasing attention is being paid to the possibility of researchers and decision-makers working together to design and subsequently evaluate interventions. One such example is the Practical Approach to Lung Health in South Africa (PALSA) trial, conducted in South Africa.[34] From the start, researchers worked with front-line clinicians and managers to develop a guideline for the management of respiratory disease, which was tested and refined prior to implementation and evaluation. This approach is not without controversy, not least because decision-makers typically have a commitment to making an intervention work, which may impact on the evaluation and interpretation. For example, a metaregression describing features of effective computerised clinical decision support systems demonstrated that studies conducted by the system developers were more likely to show benefit than those conducted by a third party.[35]

Importance of process evaluation

Randomised controlled trials remain the best method for making causal inference and providing a reliable basis for decision-making. However, the basic RCT design cannot assess how or why an intervention does or does not achieve outcomes. To understand this, process evaluations should be routinely included in all evaluations of complex interventions to explain discrepancies between expected and observed outcomes, to understand how context influences outcomes and to provide insights to aid further implementation.[1]

Process evaluation can usefully investigate how the intervention was delivered, providing decision-makers with information about how it might be replicated. Issues considered may include training and support, communication and management structures, and how these structures interact with attitudes and circumstances of staff to shape the intervention. Process evaluations also commonly investigate the reach of interventions, for example whether or not the intended audience comes into contact with the intervention, and how.[3] On the other hand, interventions may work, but in ways that were not expected or predicted.[34] Qualitative or quantitative process evaluation may allow modification or updating of 'logic models' to account for the new data and plan further research. Further consideration is given to the role of qualitative methods in evaluative research in *Essay 7*.

The recently published MRC guidance on process evaluation of complex interventions provides a general framework for conducting and reporting this type of study.[3] Grant *et al.*[36] also provide a framework for design and reporting which is specific to cRCTs and encourages researchers to consider both the processes involving individuals in the target population on whom outcome data are collected and the processes relating to the clusters within which individuals are nested.

'Realist' evaluation methods purport to go beyond conventional evaluations to consider not just 'what works', but 'what works, for whom, and in what circumstances?' In order to answer these questions, realist evaluators aim to develop and continually update relevant programme theory, to identify the underlying generative mechanisms that explain 'how' the outcomes were caused and the influence of context.

Although realist methods are often contrasted with conventional RCTs and systematic reviews, the differences may not be so acute.[2] For example, quantitative methods linked to RCTs exist for the analysis of mediation (to explore 'how') and moderation (to assess 'for whom'), which can usefully complement qualitative work.[37] Bonell *et al.*[2] also propose that 'realist' aims can be examined in the context of RCTs. Such 'realist' RCTs should aim to examine the effects of intervention components separately and in combination, for example using multiarm studies and factorial designs; explore mechanisms of change, for example how pathway variables mediate intervention effects; use multiple RCTs to test how intervention effects vary with context; draw on complementary qualitative and quantitative data; and be oriented towards building and validating 'mid-level' programme theories which would set out how interventions interact with context to produce outcomes.[2] 'Realist' RCTs, which not only provide an unbiased average effect estimate but also test and refine programme theory, are a potentially important extension of existing pragmatic RCT methods. Examples where this approach has been used successfully include a cRCT assessing the impact of school fruit 'tuck shops' on children's consumption of fruit and snacks in deprived areas in south Wales and South West England[38] and an evaluation of a peer-led intervention that aimed to prevent smoking uptake in secondary schools in England and Wales.[39]

The addition of process evaluations to RCTs does highlight potential tensions between quality and depth of evaluation, and its speed and efficiency. Process evaluations may require detailed qualitative and observational work, which may add significantly to the costs of the research and its duration. There is also a tension between an 'independent' process evaluation designed to make sense of the eventual trial results, and the desire for feedback from the evaluation mid-stream to deal with issues that may be limiting effectiveness.[3]

Designing for implementation

To successfully improve health care and population health, we need complex interventions with a degree of flexibility, resilient to contextual variation and therefore more likely to be implemented widely. We also require a clear understanding of contextual dependencies (such as target group, resources, system requirements) and system impacts. As we have outlined, attention to implementation issues in intervention design and understanding how or why an intervention works in the trial setting are both crucial to determining whether or not the intervention can be reliably disseminated to other settings.

The traditional linear 'phases of research' model remains influential. However, this may not be optimal for complex interventions, because it assumes that real-world effectiveness research naturally follows from successful efficacy research. As we outlined at the start of this essay, complex interventions interact with the context in which they are delivered. Contextual factors that may moderate the effectiveness of the intervention across settings and populations may not be adequately considered in design, which means that an intervention proven to be efficacious in one context may fail to demonstrate effectiveness when implemented in contexts for which it has not been designed.

Work in implementation science seeks to challenge the linear model of translation. It has been suggested that the characteristics which cause an intervention to be successful in efficacy research (e.g. intensive, complex, highly standardised) are fundamentally different from, and often at odds with, programmes that succeed in population-based effectiveness settings (e.g. having broad appeal, being adaptable for both participants and intervention agents).[40] The Reach, Effectiveness, Adoption, Implementation, Maintenance (RE-AIM) model proposed for evaluating complex interventions is intended to refocus priorities on wider issues, and give balanced emphasis to internal and external validity.[41] The model seeks to determine the characteristics of interventions that can reach large numbers of people, especially those who can most benefit; be widely adopted by different settings; be consistently implemented by staff with moderate levels of training and expertise; and produce replicable and long-lasting effects (and minimal negative impacts) at reasonable cost. An alternative model, normalisation process theory, highlights a series of constructs representing different types of work that people do in implementation that could inform intervention design, and linking it to ideas about agency, social and cognitive mechanisms, and collective action from sociology and psychology.[42] A range of other relevant frameworks and models have been reviewed.[43,44] For further discussion of current issues in implementation, see *Essay 8* in this volume.

Greater attention needs to be paid to planning and documenting intervention reach, adoption, implementation and maintenance. One solution is to extend feasibility and pilot studies. Such studies would not only optimise evaluation methods, and further develop underlying theories or logic models, but also allow detailed consideration of how interventions would be adopted, implemented and maintained in real-world contexts. However, such methods may increase the cost and duration of evaluation and may not always be feasible.

Trial speed and efficiency

As noted in *Selecting a randomised approach*, a key limitation of current RCTs is their lack of speed and high cost, and there is significant interest in ways of making RCT delivery more efficient.

Randomised controlled trials rely on patients, staff and organisations to volunteer to take part. However, patient recruitment and retention remains a significant problem[45] and the evidence base for recruitment and retention is very sparse.[46,47] Many health organisations are disengaged with the research production infrastructure, which has implications for the generalisability of findings. To maximise replicability, studies need to be conducted in a wide range of settings to avoid the clear gaps between the populations in research and those to which the results are expected to apply.

There is an urgent need for research to support the more efficient delivery of RCTs.[48] This might involve use of qualitative research to explore delivery[49] or patient and public involvement to design trials more sensitive to the needs of patients and other stakeholders.[50] More consistent outcome measurement and reporting[51] might avoid excessive measurement burden. Methodological studies may be usefully embedded in existing trials to grow the evidence base for efficient delivery.[52,53] The cohort multiple RCT uses routine measurement of patients in a cohort to facilitate both recruitment and outcome measurement. Routine measurement of patients in the cohort eases the identification of patients for RCTs, while the cohort may provide a recruitment process which more naturally replicates clinical practice.

There are exemplars for more efficient trial delivery. The Ontario Printed Education Message Trials demonstrate the possibility of conducting large-scale, randomised studies at relatively low cost.[54] General practitioners in Ontario, Canada, were randomised by practice to receive a range of reminders about retinal screening in diabetic patients. The reminders were mailed to each physician in conjunction with a widely read professional newsletter. All general practitioners in the province were included in the study, and hence the result provides both a pragmatic and a reliable assessment of the intervention in that context. Alongside a study of the effectiveness of the various interventions,[54] the authors also conducted a theory-based process evaluation to assess physicians' intentions related to referring patients for retinopathy screening, as well as their attitude, subjective norm and perceived behavioural control.[55]

Medical care and associated research is undergoing something of a paradigm shift with the introduction and use of electronic health records. These provide major opportunities owing to their scope and amenability to analysis. The idea of a 'learning health-care system' is based in part on better use of these data to support rapid evaluation (including RCTs and other quasi-experiments) and system-wide learning.[56] The potential power of electronic health records in the context of RCTs is beginning to receive attention, particularly the potential to study large samples at relatively low cost, driving improvements in efficiency.[57] In the USA, Clinical Data Research Networks supported by PCORnet (National Patient-Centered Clinical Research Network) build capacity for conducting both RCTs and observational effectiveness studies by collating data in a compatible format from a range of different health-care providers, including integrated delivery systems, academic medical centres and safety net clinics.[58]

Similarly in the UK, cRCTs have been implemented using electronic health records to evaluate the effectiveness of computer-delivered decision support tools in reducing antibiotic prescribing for respiratory tract infections in primary care[59] and improving risk factor control after first stroke.[60] Both studies made use of electronic health records available via the Clinical Practice Research Datalink (CPRD), which includes anonymised data for approximately 7–8% of UK general practices, linked to a range of other data including mortality information and cancer registrations.[61] This experience provides compelling evidence that RCTs may be performed efficiently in large samples of this kind. Process evaluation may also be carried out alongside this type of study.[62]

However, there are significant challenges. Those who contribute information to data sets such as CPRD are skilled at collecting data, and not necessarily representative of their peers. The geographical dispersal of sites in RCTs using electronic health records may present a challenge for research governance and intervention implementation.[63] For example, in a cRCT which used electronic records to study antibiotic prescribing, because the practices contributing to the CPRD are anonymous, permission had to be sought from 155 separate commissioning organisations for the English general practices alone. Further permissions were required in Scotland, Wales and Northern Ireland. The process took almost a year to complete before the RCT could commence.[63] Such requirements make it difficult to realise the potential efficiency gains of such trials.

There are always challenges concerning the accuracy of records, and these records also provide a potentially partial view of outcomes. For example, such records rarely include patient-centred outcomes such as self-reported quality of life or patient satisfaction. Sometimes, the lack of patient-oriented outcomes is less of a problem (e.g. where the focus is on professional behaviour), but there is a danger that the efficiency of RCTs using electronic health records means that patient-oriented outcomes are de-emphasised compared with 'harder' measures such as mortality, or policy-maker and manager priorities such as quality of care and health-care utilisation (especially hospital admissions).

Working with decision-makers

Finally, although the purpose of evaluation is to inform decision-making by policy-makers and service providers, there is currently something of a disconnect between those who commission research and those who deliver it. As we have outlined, challenges may arise because policy-makers are often keen to pilot a new intervention in favourable circumstances, and there may be an unwillingness to evaluate at all, in case the programme is not effective. Others argue that coproduction of research with decision-makers results in a loss of researcher independence. However, there are innovative examples of attempts to bridge this gap, such as the National Institute for Health Research Collaborations for Leadership in Applied Health Research and Care in England. Evaluability assessments have also been trialled in Scotland, encouraging early dialogue between researchers and decision-makers to establish whether or not and how to evaluate programmes and policies.[64]

One of the principal challenges, however, is managing potentially conflicting time scales. Decision-makers typically require a swift answer, which may be challenging to deliver using rigorous methods. This may be compounded where funding and regulatory approvals must be secured before an evaluation can commence, which may make it difficult to conduct baseline assessments prior to an intervention. On the other hand, in health systems in repeated flux, managerial staff may not be in post for extended periods of time, making it challenging for researchers to establish long-term relationships with key decision-makers. In these circumstances, service needs may over-rule methodological concerns, and researchers may have to use evaluation methods that would provide the best possible answer in the time available.

Conclusions

There is significant demand from decision-makers for evaluations and many complex interventions and service transformations remain unevaluated. Evaluation methodology has progressed significantly in the past 10 years. cRCTs are now well established, stepped-wedge designs show great promise, and there are an increasing array of non-randomised methods available. There is increasing acceptance of the importance of theory in intervention design, and of the role of process evaluation in helping to interpret why interventions succeed and fail.

Nevertheless, we have identified important tensions which face the evaluation community. Development of interventions which are efficacious is no longer sufficient. There is now a demand for interventions which deliver better outcomes, can achieve those benefits in different contexts and can be implemented at pace and scale to ensure that local trial impacts translate into population health impact. At the same time, competition for research funding remains acute, and funders and decision-makers demand faster, more efficient evaluations that can better meet their needs and timelines. These demands will place a premium on continued methodological development to ensure that evaluation keeps pace with innovation in health services.

Acknowledgements

We thank Martin Gulliford for helpful feedback.

Contributions of authors

Helen Barratt (Senior Clinical Research Associate, Health Services Research) and **Peter Bower** (Professor, Health Services Research) wrote the first draft of the essay.

Marion Campbell (Professor, Health Services Research), **Laurence Moore** (Professor, Public Health Research) and **Merrick Zwarenstein** (Professor, Family Medicine) commented on the draft and provided additional material.

References

1. Craig P, Dieppe P, MacIntyre S, Michie S, Nazareth I, Petticrew M. *Developing and Evaluating Complex Interventions: New Guidance*. London: MRC; 2008.
2. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc Sci Med* 2012;**75**:2299–306. http://dx.doi.org/10.1016/j.socscimed.2012.08.032
3. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, *et al.* Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 2015;**350**:h1258. http://dx.doi.org/10.1136/bmj.h1258
4. Richards D, Hallberg I. *Complex Interventions in Health: An Overview of Research Methods*. Oxford: Routledge; 2015.
5. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. *Milbank Q* 2012;**90**:421–56. http://dx.doi.org/10.1111/j.1468-0009.2012.00670.x
6. Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JPA, *et al.* Biomedical research: increasing value, reducing waste. *Lancet* 2014;**383**:101–4. http://dx.doi.org/10.1016/S0140-6736(13)62329-6
7. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG, Donner A. Methods in health service research. Evaluation of health interventions at area and organisation level. *BMJ* 1999;**319**:376–9. http://dx.doi.org/10.1136/bmj.319.7206.376
8. Moberg J, Kramer M. A brief history of the cluster randomised trial design. *J R Soc Med* 2015;**108**:192–8. http://dx.doi.org/10.1177/0141076815582303
9. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* 2005;**2**:99–107. http://dx.doi.org/10.1191/1740774505cn071oa

10. Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol* 2005;**58**:246–51. http://dx.doi.org/10.1016/j.jclinepi.2004.08.012
11. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;**35**:1292–300. http://dx.doi.org/10.1093/ije/dyl129
12. Campbell MK, Thomson S, Ramsay CR, MacLennan GS, Grimshaw JM. Sample size calculator for cluster randomized trials. *Comput Biol Med* 2004;**34**:113–25. http://dx.doi.org/10.1016/S0010-4825(03)00039-8
13. Campbell MK, Elbourne DR, Altman DG, CONSORT group. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004;**328**:702–8. http://dx.doi.org/10.1136/bmj.328.7441.702
14. Campbell MK, Piaggio G, Elbourne DR, Altman DG. CONSORT 2010 statement: extension to cluster randomised trials. *BMJ* 2012;**345**:e5661. http://dx.doi.org/10.1136/bmj.e5661
15. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c332. http://dx.doi.org/10.1136/bmj.c332
16. Weijer C, Grimshaw JM, Eccles MP, McRae AD, White A, Brehaut JC, *et al.* The Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials. *PLOS Med* 2012;**9**:e1001346. http://dx.doi.org/10.1371/journal.pmed.1001346
17. Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, *et al.* Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 2015;**16**:353. http://dx.doi.org/10.1186/s13063-015-0839-2
18. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;**6**:54. http://dx.doi.org/10.1186/1471-2288-6-54
19. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;**350**:h391. http://dx.doi.org/10.1136/bmj.h391
20. Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C, *et al.* Five questions to consider before conducting a stepped wedge trial. *Trials* 2015;**16**:350. http://dx.doi.org/10.1186/s13063-015-0841-8
21. Mdege ND, Man M-S, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;**64**:936–48. http://dx.doi.org/10.1016/j.jclinepi.2010.12.003
22. Murphy S, Moore GF, Tapper K, Lynch R, Clarke R, Raisanen L, *et al.* Free healthy breakfasts in primary schools: a cluster randomised controlled trial of a policy intervention in Wales, UK. *Public Health Nutr* 2011;**14**:219–26. http://dx.doi.org/10.1017/S1368980010001886
23. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;**350**:h2147. http://dx.doi.org/10.1136/bmj.h2147
24. Semple S, Creely KS, Naji A, Miller BG, Ayres JG. Secondhand smoke levels in Scottish pubs: the effect of smoke-free legislation. *Tob Control* 2007;**16**:127–32. http://dx.doi.org/10.1136/tc.2006.018119
25. Fleming PJ, Blair PS, Bacon C, Bensley D, Smith I, Taylor E, *et al.* Environment of infants during sleep and risk of the sudden infant death syndrome: results of 1993–5 case–control study for confidential inquiry into stillbirths and deaths in infancy. Confidential enquiry into stillbirths and deaths regional coordinators and researchers. *BMJ* 1996;**313**:191–5. http://dx.doi.org/10.1136/bmj.313.7051.191

26. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;**367**:1821–8. http://dx.doi.org/10.1056/NEJMsa1114951

27. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;**17**:2265–81. http://dx.doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

28. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6409. http://dx.doi.org/10.1136/bmj.f6409

29. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;**29**:722–9. http://dx.doi.org/10.1093/ije/29.4.722

30. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, *et al.* Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health* 2012;**66**:1182–6. http://dx.doi.org/10.1136/jech-2011-200375

31. National Institute for Health Research CLAHRC Greater Manchester. *NHS Greater Manchester Primary Care Demonstrator Evaluation: Final Report*. NIHR CLAHRC Greater Manchester; 2015. URL: http://clahrc-gm.nihr.ac.uk/wp-content/uploads/PCDE-final-report-full-final.pdf (accessed 21 April 2016).

32. Abdalrahman B, Soljak M. NHS health checks: an update on the debate and program implementation in England. *J Ambul Care Manage* 2015;**38**:5–9. http://dx.doi.org/10.1097/JAC.0000000000000070

33. Albarracín D, Gillette JC, Earl AN, Glasman LR, Durantini MR, Ho M-H. A test of major assumptions about behavior change: a comprehensive look at the effects of passive and active HIV-prevention interventions since the beginning of the epidemic. *Psychol Bull* 2005;**131**:856–97. http://dx.doi.org/10.1037/0033-2909.131.6.856

34. Fairall LR, Zwarenstein M, Bateman ED, Bachmann M, Lombard C, Majara BP, *et al.* Effect of educational outreach to nurses on tuberculosis case detection and primary care of respiratory illness: pragmatic cluster randomised controlled trial. *BMJ* 2005;**331**:750–4. http://dx.doi.org/10.1136/bmj.331.7519.750

35. Roshanov PS, Fernandes N, Wilczynski JM, Hemens BJ, You JJ, Handler SM, *et al.* Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* 2013;**346**:f657. http://dx.doi.org/10.1136/bmj.f657

36. Grant A, Treweek S, Dreischulte T, Foy R, Guthrie B. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. *Trials* 2013;**14**:15. http://dx.doi.org/10.1186/1745-6215-14-15

37. Emsley R, Dunn G, White IR. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Stat Methods Med Res* 2010;**19**:237–70. http://dx.doi.org/10.1177/0962280209105014

38. Moore L, Tapper K. The impact of school fruit tuck shops and school food policies on children's fruit consumption: a cluster randomised trial of schools in deprived areas. *J Epidemiol Community Health* 2008;**62**:926–31. http://dx.doi.org/10.1136/jech.2007.070953

39. Campbell R, Starkey F, Holliday J, Audrey S, Bloor M, Parry-Langdon N, *et al.* An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *Lancet* 2008;**371**:1595–602. http://dx.doi.org/10.1016/S0140-6736(08)60692-3

40. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *Am J Public Health* 2003;**93**:1261–7. http://dx.doi.org/10.2105/AJPH.93.8.1261

41. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health* 1999;**89**:1322–7. http://dx.doi.org/10.2105/AJPH.89.9.1322

42. May C. Towards a general theory of implementation. *Implement Sci* 2013;**8**:18. http://dx.doi.org/10.1186/1748-5908-8-18

43. Tabak RG, Khoong EC, Chambers D, Brownson RC. Bridging research and practice. *Am J Prevent Med* 2012;**43**:337–50. http://dx.doi.org/10.1016/j.amepre.2012.05.024

44. Moullin JC, Sabater-Hernández D, Fernandez-Llimos F, Benrimoj SI. A systematic review of implementation frameworks of innovations in healthcare and resulting generic implementation framework. *Health Res Policy Syst* 2015;**13**:16. http://dx.doi.org/10.1186/s12961-015-0005-z

45. Sully BGO, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. *Trials* 2013;**14**:166. http://dx.doi.org/10.1186/1745-6215-14-166

46. Brueton VC, Tierney JF, Stenning S, Meredith S, Harding S, Nazareth I, *et al.* Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. *BMJ Open* 2014;**4**:e003821. http://dx.doi.org/10.1136/bmjopen-2013-003821

47. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, *et al.* Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. *BMJ Open* 2013;**3**:e002360. http://dx.doi.org/10.1136/bmjopen-2012-002360

48. Treweek S, Altman DG, Bower P, Campbell M, Chalmers I, Cotton S, *et al.* Making randomised trials more efficient: report of the first meeting to discuss the Trial Forge platform. *Trials* 2015;**16**:261. http://dx.doi.org/10.1186/s13063-015-0776-0

49. Donovan JL, Paramasivan S, de Salis I, Toerien M. Clear obstacles and hidden challenges: understanding recruiter perspectives in six pragmatic randomised controlled trials. *Trials* 2014;**15**:5. http://dx.doi.org/10.1186/1745-6215-15-5

50. Ennis L, Wykes T. Impact of patient involvement in mental health research: longitudinal study. *Br J Psychiatry* 2013;**203**:381–6. http://dx.doi.org/10.1192/bjp.bp.112.119818

51. Gargon E, Gurung B, Medley N, Altman DG, Blazeby JM, Clarke M, *et al.* Choosing important health outcomes for comparative effectiveness research: a systematic review. *PLOS ONE* 2014;**9**:e99111. http://dx.doi.org/10.1371/journal.pone.0099111

52. Rick J, Graffy J, Knapp P, Small N, Collier DJ, Eldridge S, *et al.* Systematic techniques for assisting recruitment to trials (START): study protocol for embedded, randomized controlled trials. *Trials* 2014;**15**:407. http://dx.doi.org/10.1186/1745-6215-15-407

53. Smith V, Clarke M, Devane D, Begley C, Shorter G, Maguire L. SWAT 1: what effects do site visits by the principal investigator have on recruitment in a multicentre randomized trial? *J Evid Based Med* 2013;**6**:136–7. http://dx.doi.org/10.1111/jebm.12049

54. Zwarenstein M, Shiller SK, Croxford R, Grimshaw JM, Kelsall D, Paterson JM, *et al.* Printed educational messages aimed at family practitioners fail to increase retinal screening among their patients with diabetes: a pragmatic cluster randomized controlled trial [ISRCTN72772651]. *Implement Sci* 2014;**9**:87. http://dx.doi.org/10.1186/1748-5908-9-87

55. Grimshaw JM, Presseau J, Tetroe J, Eccles MP, Francis JJ, Godin G, *et al.* Looking inside the black box: results of a theory-based process evaluation exploring the results of a randomized controlled

trial of printed educational messages to increase primary care physicians' diabetic retinopathy referrals [Trial registration number ISRCTN72772651]. *Implement Sci* 2014;**9**:86. http://dx.doi.org/10.1186/1748-5908-9-86

56. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff (Millwood)* 2014;**33**:1163–70. http://dx.doi.org/10.1377/hlthaff.2014.0053

57. Van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, *et al.* The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;**18**(43). http://dx.doi.org/10.3310/hta18430

58. Clinical Data Research Networks. *PCORnet*. 2016. URL: www.pcornet.org/clinical-data-research-networks/ (accessed February 2016).

59. Gulliford MC, Staa T van, Dregan A, McDermott L, McCann G, Ashworth M, *et al.* Electronic health records for intervention research: a cluster randomized trial to reduce antibiotic prescribing in primary care (eCRT study). *Ann Fam Med* 2014;**12**:344–51. http://dx.doi.org/10.1370/afm.1659

60. Dregan A, van Staa TP, McDermott L, McCann G, Ashworth M, Charlton J, *et al.* Point-of-care cluster randomized trial in stroke secondary prevention using electronic health records. *Stroke* 2014;**45**:2066–71. http://dx.doi.org/10.1161/STROKEAHA.114.005713

61. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, *et al.* Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015;**44**:827–36. http://dx.doi.org/10.1093/ije/dyv098

62. McDermott L, Yardley L, Little P, van Staa T, Dregan A, McCann G, *et al.* Process evaluation of a point-of-care cluster randomised trial using a computer-delivered intervention to reduce antibiotic prescribing in primary care. *BMC Health Serv Res* 2014;**14**:594. http://dx.doi.org/10.1186/s12913-014-0594-1

63. Gulliford MC, van Staa TP, McDermott L, McCann G, Charlton J, Dregan A, *et al.* Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT study). *Trials* 2014;**15**:220. http://dx.doi.org/10.1186/1745-6215-15-220

64. Craig P, Campbell M. *Evaluability Assessment: A Systematic Approach to Deciding Whether and How to Evaluate Programmes and Policies (Working Paper)*. What Works Scotland; 2015. URL: http://whatworksscotland.ac.uk/wp-content/uploads/2015/07/WWS-Evaluability-Assessment-Working-paper-final-June-2015.pdf (accessed February 2016).

Essay 3 Advancing quantitative methods for the evaluation of complex interventions

Clare Gillies,[1] Nick Freemantle,[2] Richard Grieve,[3]
Jasjeet Sekhon[4] and Julien Forder[5]

[1]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) East Midlands and NIHR Research Design Service East Midlands, University of Leicester, Leicester, UK
[2]Department of Primary Care and Population Health, University College London, London, UK
[3]Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK
[4]Department of Political Science and Statistics, University of California Berkeley, Berkeley, CA, USA
[5]School of Social Policy, Sociology and Social Research, University of Kent, Canterbury, UK

Declared competing interests of authors: none

Published May 2016
DOI: 10.3310/hsdr04160-37

This essay should be referenced as follows:

Gillies C, Freemantle N, Grieve R, Sekhon J, Forder J. Advancing quantitative methods for the evaluation of complex interventions. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 37–54.

List of figures

List of boxes

List of abbreviations

AQ	Advance Quality
ASCOT	Adult Social Care Outcomes Toolkit
CI	confidence interval
IV	instrumental variable
PHB	personal health budget
RALES	Randomised Aldactone Evaluation Study
RCT	randomised controlled trial
SCRQOL	social care-related quality of life
WSD	Whole Systems Demonstrator

Abstract

An understanding of the impact of health and care interventions and policy is essential for decisions about which to fund. In this essay we discuss quantitative approaches in providing evaluative evidence. *Experimental* approaches allow the use of 'gold-standard' methods such as randomised controlled trials to produce results with high internal validity. However, the findings may be limited with regard to generalisation: that is, feature reduced externality validity. *Observational* quantitative approaches, including matching, synthetic control and instrumental variables, use administrative, survey and other forms of 'observational' data, and produce results with good generalisability. These methods have been developing in the literature and are better able to address core challenges such as selection bias, and so improve internal validity. Evaluators have a range of quantitative methods available, both experimental and observational. It is perhaps a combination of these approaches that is most suited to evaluating complex interventions.

Scientific summary

An understanding of the impact of health and care interventions and policy is essential to inform decisions about which to fund. Quantitative approaches can provide robust evaluative evidence about the causal effects of intervention choices.

Randomised controlled trials (RCTs) are well established. They have good internal validity: that is, they produce accurate estimates of causal effects for the study participants, minimising selection bias (confounding by indication). The findings may, however, be limited with regard to generalisation: that is, feature reduced externality validity.

Observational quantitative approaches, which use data on actual practice, can produce results with good generalisability. These methods have been developing in the literature and are better able to address core challenges such as selection bias, and so improve internal validity.

This essay aims to summarise a range of established and new approaches, discussing the implications for improving internal and external validity in evaluations of complex interventions.

Randomised controlled trials can provide unbiased estimates of the relative effectiveness of different interventions within the study sample. However, treatment protocols and interventions can differ from those used in routine practice, and this can limit the generalisability of RCT results. To address this issue, trial samples can be reweighted using observational data about the characteristics of people in routine practice, comparing the outcomes of people in the trial with those in practice settings. Evidence for similarity of outcomes can be assessed using 'placebo tests'.

Observational studies may provide effect estimates confounded by indication (i.e. exhibit treatment-selection bias) because the factors that determine actual treatment options for individuals are also likely to affect their treatment outcomes. Observational studies seek to address selection by trying to remove the consequences of the selection process. This can be done by using data on all relevant selection factors, applying matching methods, including recently developed 'genetic' matching, or using regression control.

When selection is likely to be influence by unobserved factors, alternative methods are available that exploit the existence of particular circumstances that structure the problem and data. These include instrumental variables, regression discontinuity and the difference-in-difference.

There a growing need to demonstrate effectiveness and cost-effectiveness of complex interventions. This essay has shown that evaluators have a range of quantitative methods, both experimental and observational. However, it is perhaps the use of a combination of these approaches that might be most suited to evaluating complex interventions.

Introduction

An understanding of the impact of different health and care interventions is essential in helping us to make the best choices when deciding which interventions and treatments to fund. Quantitative approaches can provide robust evaluative evidence about the causal effects of intervention choices. Randomised controlled trials (RCTs) are well established, but recently there have been significant advances and refinements in observational approaches, allowing complex interventions to be assessed in more pragmatic settings.

There are significant challenges in the evaluation of complex interventions. Complexity can be taken to mean *complicated* in the sense of interventions with many interdependency components but, more pertinently, complexity can also refer to systems that adapt to context and exhibit non-linear responses.[1]

Many health and care policy ‘interventions’ can be regarded as complex in the latter sense, for example policies to create closer integration across the care system or to support person-centred approaches to management of chronic disease.

Observational or non-experimental methods, though having their own limitations, are an alternative way to produce estimates of the causal effects of complex interventions. They rely on ‘natural’ variation in a population regarding the characteristics of the intervention, and the factors that affect its outcomes. As with experimental approaches such as RCTs, establishing the counterfactual outcome is the key to evaluation. In general, treatment choice is according to anticipated prognosis or performance. Hence a study that makes unadjusted comparisons of the outcomes of those units (e.g. patients or hospitals) is liable to provide estimates of the effect of treatment that are biased owing to selection into treatment (also known as confounding by indication). On their own, the observed outcomes of non-recipients of the intervention will not be good indicators of the counterfactual outcomes that would have been experienced by people who did use the intervention.

Observational approaches seek to address this problem by trying to identify and account for the consequences of this selection process. In RCTs, people are randomly assigned between intervention and (counterfactual) control groups. In most RCTs there is a selection process for including study participants, for example according to eligibility criteria, which may include requiring informed consent. By contrast, in observational studies the inclusion criteria tend to be less stringent, which allows these studies to observe outcomes for those patients and treatments that are relevant for routine practice.

These contrasting features of RCTs and observational approaches give rise to different properties when they are used in evaluations. Two important properties are those of internal and external validity. The former, that of *internal validity*, concerns the extent to which the contrast we are making allows accurate estimates of causal effects for the study participants: that is, the study avoids selection bias due to treatment selection (confounding by indication). The second, that of *external validity*, is about the extent to which the results of a study are generalisable and applicable to routine practice: that is, the study avoids selection bias due to *sample* selection.

Randomised controlled trials have far greater internal validity owing to the randomisation process which removes the effects of selection into treatment, but often have lower external validity owing to the restrictions an experimental study places on how an intervention can be delivered and who can be included: that is, sample selection. Other essays in this volume delineate the repertoire of randomised designs now available in health services research (e.g. *Essay 2*).

Observational studies, on the other hand, should provide results that are representative of what may be achieved in practice, and therefore may have greater external validity. However, the internal validity of observational studies is compromised because the assignment of the intervention is conditional on certain (unobserved) characteristics (not the play of chance) that also affect outcomes, and this introduces bias due to treatment selection.

In other words, we would expect RCTs to provide unbiased estimates of *sample average treatment effects*, but potentially biased estimates of the treatment effects of people receiving the intervention in routine practice: that is, of the *population average treatment effects of the treated*. *Box 3.1* provides details.

When evaluating *complex* interventions, both of these validity requirements can be hard to achieve. With regard to internal validity, there are particular challenges in identifying and distinguishing the ‘intervention’, and precluding cross-contamination between intervention and control groups. Turning to external validity, because complex interventions tend to be highly context-specific in their effects, generalising the results for policy and practice requires more nuanced analyses of why effects occur, not just an estimate of the magnitude of the effect within the RCT setting.

BOX 3.1 Estimation of treatment effects

Suppose we are trying to establish the effects of some intervention. Let y^1 be the average observed outcome of people who received the intervention in question and y^0 be the average observed outcome of people who received the comparator intervention. Outcomes will be affected by (a) which people get allocated into the intervention and control/comparator groups in a study; and (b) which people are eligible and are sampled into the study. We can denote these two processes as the treatment group allocation, *t*, which can be either the intervention ($t = T$) or the control group ($t = C$), and the study eligibility group process, *s*, which, for exposition, can be either the study sample group ($s = S$) or the (real-world) population of potential recipients group ($s = P$). Expected outcomes in a group are therefore conditional on these processes: $y^1_{ts} = E[y^1|t, s]$ for people getting the intervention and $y^0_{ts} = E[y^0|t, s]$ for people not getting the intervention.

Evaluators are often interested in population average treatment effects and, more specifically, the outcome of people who would actually be assigned to the intervention in practice (i.e. in the 'real world'). This is the PATT: $PATT = y^1_{TP} - y^0_{TP}$. By contrast, the sample average treatment effect of the treated is: $SATT = y^1_{TS} - y^0_{TS}$.

In a RCT, the observed outcome from the intervention in the treatment group is $y^1_{RCT} = y^1_{TS}$ and in the control group is $y^0_{RCT} = y^0_{CS}$. As a result of randomisation, the characteristics of people in the control group will differ only on the basis of chance with those in the treatment group, i.e. $y^0_{CS} = y^0_{TS}$. Therefore, RCTs give unbiased (internally valid) estimates of SATT: $y^1_{RCT} - y^0_{RCT} = y^1_{TS} - y^0_{CS} = y^1_{TS} - y^0_{TS} = SATT$. (Random assignment means that in expectation SATT equals the sample average treatment effect.) However, because of the study eligibility criteria, RCTs might not give unbiased (externally valid) estimations of PATT. Suppose that $y^1_{TS} = y^1_{TP} + \varepsilon^1_P$ and $y^0_{CS} = y^0_{CP} + \varepsilon^0_{CP}$ where the ε terms are the differences in outcomes between the sample and population. Then $y^1_{RCT} - y^0_{RCT} = y^1_{TP} - y^0_{TP} + \varepsilon^1_{TP} - \varepsilon^0_{CP} = PATT + \varepsilon^1_{TP} - \varepsilon^0_{CP}$. This is the form of bias known as sample selection bias.

An alternative is the non-randomised study. Suppose it is based on a representative sample of actual practice: in this case *y*, $y^1_{NRS} = y^1_{TP}$ and $y^0_{NRS} = y^0_{CP}$. Without randomisation we cannot be confident that the treatment and control group have the same (average) characteristics. Therefore, $y^0_{NRS} = y^0_{CP} = y^0_{TP} - \mu^0_{TP}$, where μ^0_{TP} denotes the difference in outcomes. As a result we have a potentially biased estimate of PATT, i.e. $y^1_{NRS} - y^0_{NRS} = y^1_{TP} - y^0_{TP} + \mu^0_{TP} = PATT + \mu^0_{TP}$. The term μ^0_{TP} is generally known as bias due to treatment selection (confounding by indication).

PATT, population average treatment effect of the treated; SATT, sample average treatment effect of the treated.

This essay aims to summarise a range of both established and new approaches aimed at assessing and improving internal and external validity for both observational and experimental evaluations of complex interventions.

Improving external validity of randomised controlled trials

Why external validity is an issue

Randomised controlled trials, when carried out to a high standard, can provide unbiased estimates of the relative effectiveness of different interventions within the study sample. Although much time and attention have been given to ensuring maximisation of internal validity through high-quality design and conduct of RCTs, problems of external validity have been less rigorously addressed. In RCTs, treatment protocols and interventions can differ from those used in routine practice, and therefore results from a RCT may be unrepresentative of what would happen in practice. Moreover, regarding complex interventions, the effects of the intervention might depend closely on factors outside the study, making the results highly context specific. For example, the impact of telehealth could also be affected by local policies regarding health and social care integration.

Assumptions are being made when directly generalising the results of a RCT to target populations: first, that the RCT participants have similar characteristics to the target population and that the control intervention in the RCT equates to what is provided as usual care in routine practice; and, second, that the intervention in the RCT is delivered as it would be if rolled out in practice.

Using observational data to improve external validity of randomised controlled trials

There have been recent developments in using observational approaches and data to address the issue of external validity. These involve comparing the characteristics of RCT study samples with data from the population of people using (or eligible for) the intervention. Trial samples can be reweighted in line with the characteristics of people in routine practice, and then the outcomes of the people in the trial can be compared with those of people in routine practice. Placebo tests can then be used to assess the evidence for similarity of outcomes.

Recent work by Hartman *et al.*[2] has built on previous approaches which have considered the external validity and generalisability of results from RCTs. The work illustrates how to extrapolate from the sample average treatment effect estimated from a RCT to a population average treatment effect for the population of interest, by combining results from experimental studies with observational studies. In particular, they specify the assumptions that would be required for a reweighting of the treatment group population of a RCT – using observational data on a target population of treated individuals – to produce population-treatment effects of the treated.

As well as formally defining the assumptions required for estimating population treatment effects from RCT data, Hartman *et al.*[2] also recommend that future randomised trials should consider how the results may be combined with observational data, and consider this when designing the trial.

Placebo tests

Placebo tests are an approach for assessing model validity by testing for an effect where none should exist. Placebo tests can be used to assess whether or not the assumptions required for generalising RCTs to routine practice are likely to be met. In Hartman *et al.*,[2] the placebo tests contrast the outcomes for patients who receive the treatment in routine practice with those who receive that same treatment within the trial setting, after adjusting for differences in observed patient characteristics between the RCT and the routine practice settings. The placebo tests are formulated such that the null hypothesis is that the adjusted outcomes from the treatment group in the RCT are 'not equivalent' to the outcomes following treatment in routine practice.[2] If the null is not rejected, then this is an indication that the results of the RCT are not generalisable because of differences in the patients or the treatments between the RCT and routine practice settings, or that there is insufficient statistical power to reject the null hypothesis (with reference to *Box 3.1*, this is essentially a test of $\hat{y}^1_{RCT}(W_{TP}) + y^1_{TP}(W_{TP})$ where W_{TP} are the observed characteristics of the treated population and $\hat{y}^1_{RCT}(W_{TP})$ are the reweighted RCT outcomes for the treatment group). Placebo tests can also be used to compare control groups in a similar way between sample and target populations.

As an example, the use of placebo tests can be discussed using the Whole Systems Demonstrator (WSD) cluster randomised trial as an example.[3] This was a large RCT that evaluated telehealth against standard care. The trial randomised 3230 adults with chronic conditions to either telehealth or usual care, where the telehealth arm received home-based technology to record medical information, such as blood glucose level and weight, and to answer symptom questions. Information from patients was then transmitted automatically to monitoring centres staffed by employees from local health-care organisations. Published results were cautiously optimistic and suggested that telehealth patients experienced fewer emergency hospital admissions than controls over 12 months [incidence rate ratio 0.81, 95% confidence interval (CI) 0.65 to 1.00; $p = 0.046$].[4]

Concerns about the generalisability of the WSD trial were raised for several reasons, but in particular because emergency admission rates increased among control patients shortly after their recruitment, suggesting that these patients may not have received usual care. It was deemed unlikely that this increase in admissions represented a normal evolution in service use, and instead it was suggested that health-care professionals may have identified unmet needs while recruiting patients to the control group and changed the management of this group from usual care or, alternatively, the trial recruitment processes might have led to changes in behaviour among patients.[4]

To assess whether or not the control group in the WSD RCT was representative of usual care, placebo tests were carried out.[4] A target population of patients who met the RCT inclusion criteria and who received usual care (n = 88,830) was identified from observational data. Of these, 1293 individuals were matched with the RCT controls on 65 identified baseline covariates (e.g. demographics, blood pressure, medications). The placebo test then contrasted outcomes from the RCT between the matched target control population and the control group of the WSD trial. This process is described in greater detail in Steventon *et al.*[4]

A comparison of the RCT control arm versus the matched controls from the observational data gave an incidence rate ratio of 1.22 (95% CI 1.05 to 1.43) for emergency admissions. In other words, emergency admissions were significantly higher in the control arm of the WSD trial than in a similar group of patients who did not participate in the trial. Consequently, the trial results may have shown an inflated beneficial effect of telehealth interventions. This example highlights the importance of assessing the generalisability of RCT results to a target population. For those settings in which placebo tests fail, Steventon *et al.*[5] propose sensitivity analyses.

Improving internal validity of observational studies

Addressing the selection problem

Observational studies, no matter how large, may provide effectiveness estimates confounded by indication: that is, exhibit treatment-selection bias. Where treatment or intervention assignment uses a decision rule that accounts for the characteristics of the individual – for example, the treatment may be more likely to be given to the patients in the poorest health – observed outcomes will provide a biased estimate of the intervention if these characteristics also directly impact on outcomes. In such an example, the intervention group will have lower observed outcomes as a result of their poor health, irrespective of the effects of the treatment.

Observational studies seek to address this selection problem by trying to remove the consequences of the selection process. The basis of this approach is to account as far as possible for the difference in characteristics between people who are getting the intervention and those who are not. This method generally relies on being able to observe all relevant characteristics, which in practice is never entirely possible. Where we anticipate having data on all relevant confounders, adjustment can be made using a number of well-documented methods, including regression and matching (e.g. propensity and prognostic scoring). More recent developments in this regard include 'genetic' matching.

When we do not anticipate being able to identify all confounders, adjustment is more complex. Where selection is on unobserved characteristics, methods include the use of instrumental variables (IVs), regression discontinuity and the difference-in-difference approach. These methods exploit the existence of particular circumstances that structure the problem and data to address the selection problem. There have been a number of significant refinements of these approaches in recent years, including the use of synthetic controls and improved diagnostics.

The difference-in-difference method

The difference-in-difference approach addresses the selection problem by comparing the experiences of a control group with the intervention group before and after the intervention. The idea is that, under certain assumptions, selection bias can be controlled for by removing any difference in the outcome indicator between groups before the intervention from differences after the intervention.

From *Figure 3.1*, β_4 is the unadjusted difference in outcome indicators between the two groups after the intervention. By contrast, β_3 is the adjusted difference, and provides a (less) biased estimate of the intervention effect as it takes into account differences in the outcome measure between the two groups that were present at the start of the evaluation period. In this linear example, β_1 is the size of the selection bias. The unbiased estimate β_3 in this example is found by subtracting the difference β_1 at baseline (time t = 0) from the difference β_4 at time t = 1.

The difference-in-difference approach was utilised in an evaluation of personal health budgets (PHBs), which was a policy piloted in 2009 whereby patients were given control over their own budgets to meet their care needs.[6] A fully RCT proved unfeasible, so a pragmatic design was chosen whereby some sites were randomised while others selected participants. The study comprised 2000 participants covering a range of health conditions. For illustration, one of the main outcome measures was care-related quality of life (Adult Social Care Outcomes Toolkit; ASCOT). At baseline the PHB group had (significantly) lower ASCOT scores than the control group. When measured at follow-up, the PHB group still had lower ASCOT scores than the control group, but the gap had closed. Subtracting the larger negative difference at baseline in the scores produced a result whereby the PHB group had significantly higher (better) scores than the control group at follow-up. Without accounting for the poorer quality of life of the PHB group prior to the intervention, this selection bias might have led to opposite conclusions about the effectiveness of PHBs.

A range of assumptions is required for the difference-in-difference approach to give unbiased estimates. It can be applied where there is only a partial uptake to a new intervention, allowing for a control group, and where there is a defined before-and-after time, recorded in the data. Although it can also control for persistent differences between groups that are not related to the intervention, it is a less appropriate

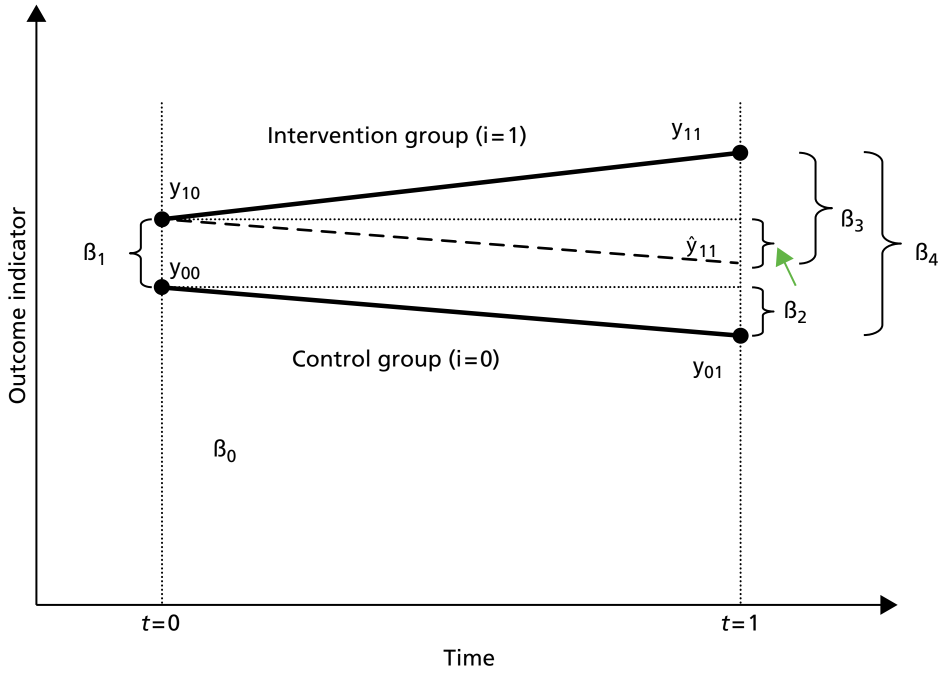

FIGURE 3.1 The difference-in-difference approach.

method where baseline factors may influence the rate of change of outcomes during the follow-up period, that is, when there are unobserved pretreatment differences between the groups whose effects change over the follow-up period. For this, the synthetic control method may be a more appropriate approach.

Synthetic controls

The synthetic control method also aims to control for treatment selection bias by using covariates and outcomes of the intervention and control groups prior to implementation to adjust outcomes after the intervention. The method exploits the availability of multiple time points and control units (e.g. different localities) prior to baseline to try and adjust for those unobserved characteristics whose effects may differ over time. A 'synthetic' control sample is constructed by weighting together multiple control units in such a way that the expected outcomes in the pre-implementation period are similar for the control and intervention groups, in order to minimise the treatment selection bias because of unobserved factors whose effects vary over time.

The synthetic control method estimates treatment effects by using this weighted synthetic control group to represent the counterfactual outcomes for the treated group after implmentation.[7] The idea behind synthetic controls is that a combination of units (where units may be hospitals, general practices, an area or region, etc.) often provide a better comparison for the unit exposed to the intervention of interest than any single unit alone. Furthermore, as the choice of synthetic control does not require access to post-intervention outcomes, this method allows researchers to decide on study design without knowing how these decisions will affect the resulting conclusions: an important issue for reducing potential research bias in observational studies.[7]

Kreif *et al.*[8] carried out an analysis to compare and contrast both the difference-in-difference method and synthetic controls for an evaluation of the Advance Quality (AQ) programme, which is a pay-for-performance initiative linking health-care provider income to quality measures. The initiative was first introduced into 24 hospitals in the North West of England, with the other nine regions in England providing a potential comparison group of 132 hospitals. *Figure 3.2* shows the trajectory of the difference from expected mortality rates before and after the implementation of the AQ programme, in both the North West and the rest of England. For control regions, the size of the difference from the expected mortality rate is relatively constant and fluctuates around zero for the period of observation. For the North West, the mortality rate is higher than would be expected before AQ, but this improves after the introduction of the AQ programme. What is clear from the graph is that there is little evidence to support the assumption of parallel trends required for a (standard) difference-in-difference approach.

For the synthetic control method, a control group was selected from the potential pool of 132 hospitals according to pre-intervention characteristics of hospital type, size and scale, and hospital quality measures. The controls were weighted according to the pretreatment trajectory of the intervention hospitals (*Figure 3.3*) and hence the synthetic control is more representative of the counterfactual outcome for the intervention hospitals.

For this particular example, when using a synthetic control comparator it was found that the AQ initiative did not significantly reduce 28-day hospital mortality. This is in contrast to the original analysis that found a significant reduction in mortality in the AQ group. This highlights the need to consider how well analysis assumptions are met when choosing an appropriate method, and also the importance of considering alternative approaches and contrasting the results. In this example, the difference-in-difference approach appeared to be flawed in that the parallel trends assumption was poorly met. The synthetic control method should be considered as an alternative approach if parallel trends are not present, although this method does require data on a sufficiently long pretreatment period to allow for the selection and weighting of an appropriate control group. More details on this analysis and other alternatives are provided in the paper by Kreif *et al.*[8]

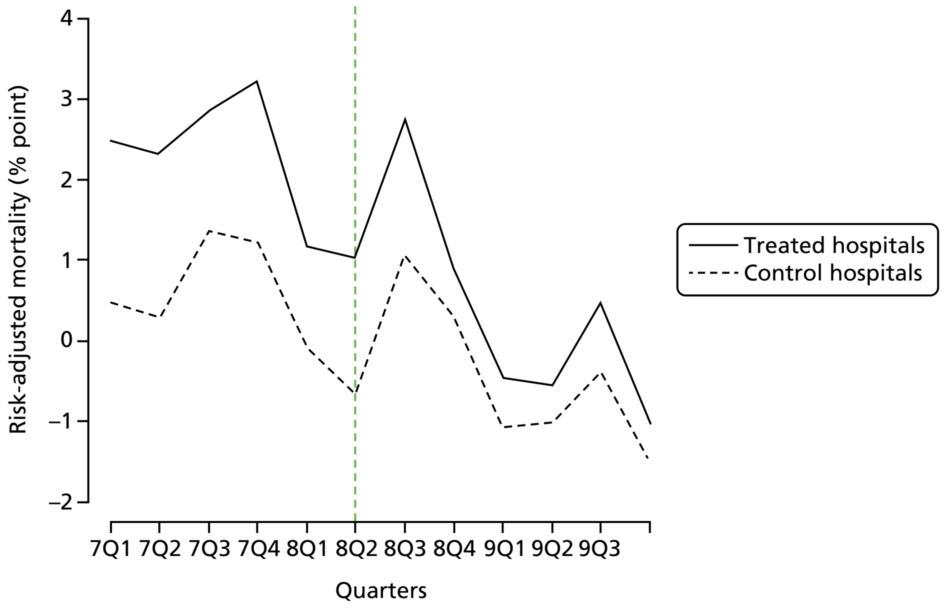

FIGURE 3.2 The observed risk adjusted mortality in pneumonia patients for the North West vs. rest of England before and after the introduction of AQ. Adapted from Kreif *et al.*,[8] © 2015, John Wiley & Sons Ltd., under the terms of the Creative Commons Attribution International Public Licence (CC BY-NC 4.0), which permits use, distribution and reproduction in any medium, provided the original work is properly cited, the use is non-commercial and is otherwise in compliance with the licence.

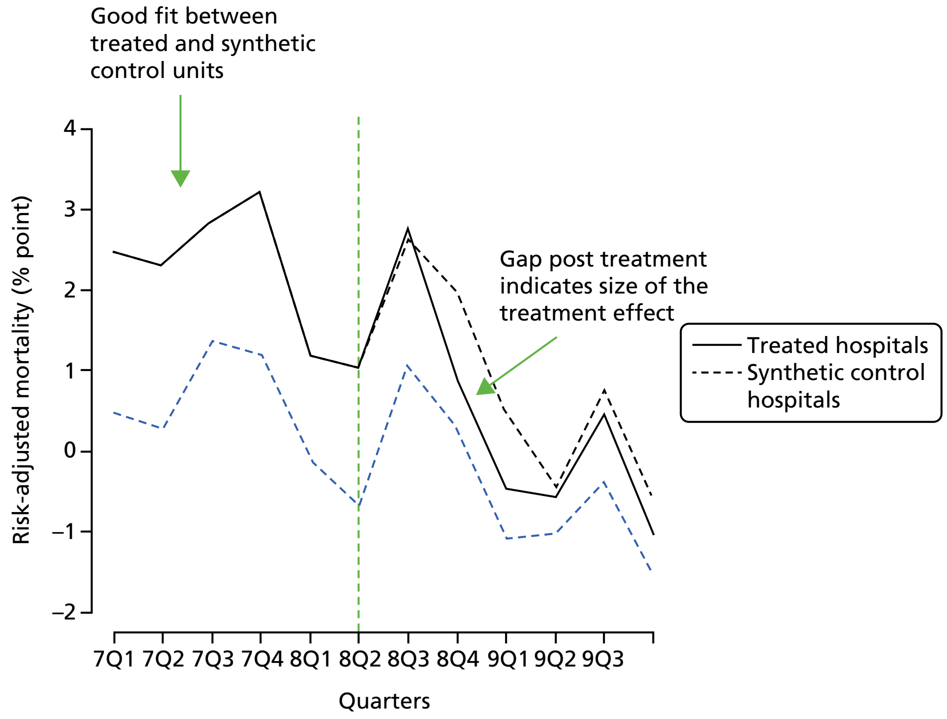

FIGURE 3.3 Comparison of the intervention hospitals with the weighted synthetic control group, pre and post intervention. Adapted from Kreif *et al.*,[8] © 2015, John Wiley & Sons Ltd., under the terms of the Creative Commons Attribution International Public Licence (CC BY-NC 4.0), which permits use, distribution and reproduction in any medium, provided the original work is properly cited, the use is non-commercial and is otherwise in compliance with the licence.

Instrumental variables

The concept of IVs has been around for over half a century, and they are widely used in economics because of the difficulty of doing controlled experiments in this field.[9] Despite their popularity in economics, they have been little used in the biostatistical literature, but their popularity is growing as they may provide a useful methodology for addressing the selection problem posed by observational studies when relevant instruments can be identified.

The essence of an IV approach is to find an indicator that is highly related to the treatment or intervention being assessed but does not independently affect the outcome of interest.[10] There is an analogy with the RCT approach in that the random allocator can be regarded as a perfect instrument.[11] In an IV analysis, the intervention variable is therefore replaced by the chosen IV, to remove the dependence of the intervention variable (or assignment) on the unobserved confounders which might influence outcomes other than through the intervention. This reduces, or in the case of very large samples eliminates, the selection bias. The use of IVs will be illustrated with an example.

A 2009 survey of service care utilisation by the elderly was carried out, with the aim of assessing the cost-effectiveness of the services provided.[12] In this case, the intensity of service utilisation was expected to be directly related to a patient's level of frailty and ill-health, as these factors are key elements in the patient's assessment of need. Therefore, methods had to be considered which distinguished the variation in care-related quality of life owing to service use as opposed to other factors.

Data were collected on a range of outcome measures, including a measure of social care-related quality of life (SCRQOL), using the ASCOT measure.[13] As is common with health and care utilisation data, the observed relationship between the intensity of service utilisation and the outcome (care-related quality of life) was negatively sloped. In other words, unadjusted data would suggest a negative effect of care utilisation on care-related quality of life. However, this observed result was probably attributable to selection bias because the amount of service a person was offered in the care system was related to their baseline level of need (e.g. severity of health condition), which was in turn negatively correlated with their quality of life. Observed confounders – for example indicators of need or poor health – can be used to control for selection, but there may be many other unobserved factors that influence both selection (i.e. in this case the amount of service) and outcomes.

To address these, an IV analysis was carried out. An IV was needed that was related to the amount of service a person received but not their current SCRQOL score. As different local authorities used different service eligibility policies, this characteristic could be used as an instrument. With the use of the IV controlling for both known and unknown confounders, a significant positive association between intensity of service use and care-related quality of life was found.

Instrumental variables have a number of applications and have been used in recent studies assessing NHS expenditure on mortality and quality of life,[14] impact of social care use on hospital utilisation[15] and impact of care services on hospital delayed discharge rates,[16] to name but a few. They can be an effective way to deal with unobservables, and are useful for adjusting for treatment selection problems that occur in non-randomised studies, but, as with all methods, there are limitations.[11,17] The key challenge is to find suitable instruments and, once chosen, to demonstrate that the choice of IV was appropriate. A crucial assumption for the correct use of IVs is that the IV is (strongly) associated with the intervention of interest but not directly to the outcome. This condition cannot be directly evaluated, although a range of diagnostic tests may help to inform instrument specification. Nonetheless, the choice of the IV is often largely based on prior judgement. Poor or 'weak' instruments will lead to significant bias, potentially offsetting reductions in selection bias. IV analyses need a large sample, and results from an IV analysis may still be biased. Recommendations and guidance around the use of IVs for the evaluation of complex interventions would be a useful step for taking these methods forward.

Propensity scores

The above sections discuss methods that are suitable when not all of the selection characteristics are known. When it is believed that all the factors on which selection of the intervention group has been based are known and observed, propensity scores can be used. Propensity scores were first described in the early 1980s.[18] A propensity score is developed by using a statistical model to estimate the individual likelihood that patients will receive treatment, using known explanatory variables such as demographic and medical history, prescribed drugs and consultation behaviour. This enables a propensity score to be calculated for each patient: that is, the likelihood that they will receive treatment or 'fitted value'. Propensity scores can then be used either for adjustment in statistical models or to create matched groups by selecting treatment and control groups with similar propensity scores.[19] Propensity score approaches have been used extensively in applied health research.

Although propensity scores go some way towards adjusting for selection bias in observational studies, they do not assure internal validity. Propensity scores can only be based on known and observed patient characteristics.

There is evidence of how propensity score approaches can fail to adequately address bias. The Randomised Aldactone Evaluation Study (RALES)[20] evaluated spironolactone (Aldactone, G.D. Searle LLC), an aldosterone inhibitor, compared with placebo, and found that it reduced mortality in patients with severe heart failure (hazard ratio 0.70, 95% CI 0.60 to 0.82; $p < 0.001$). The results of this RCT were subsequently confirmed in two further trials.[21,22]

Freemantle *et al.*[19] attempted to replicate the results of the RALES and subsequent trials using data from the Health Improvement Network, with the ultimate objective of bridging from the trial population to a real-world population of people with heart failure. As it was believed that a number of factors would influence the prescribing of spironolactone to patients, a complex propensity score was developed to account for this, which included information on patient demographics, comorbidities and current drug treatments. Two groups ($n = 4412$) treated and not treated with spironolactone, tightly matched on propensity score, were then selected for the analysis. An initial comparison of the two matched groups appeared to show results contradictory to those from the RALES trial, in that patients treated with spironolactone had an increased mortality risk (hazard ratio 1.32, 95% CI 1.18 to 1.47; $p < 0.001$).

The performance of a propensity score may be evaluated by assessing its homogeneity at different points on the scale. *Figure 3.4* shows that, for different values of the propensity score (divided into quartiles), different estimates of the effect of spironolactone on mortality risk were estimated. If the propensity score had worked well, one would expect to see similar treatment effects across all strata. For this example, it appears that the prescriber making the clinical decision to treat with spironolactone used additional information on severity of heart failure that was not captured by the propensity score. Therefore, the matching of the two groups for the analysis was not appropriate, resulting in individuals with too low a risk of mortality being matched to the spironolactone patients. In particular, this affects the results of the apparently low-propensity subjects: that is, the lowest two quartiles of the propensity score (see *Figure 3.4*), where the hazard ratio for spironolactone is particularly high, although a bias is present across the entire range of the score.

Propensity scores assume that all important variables have been included and that there are no additional processes related to disease severity that are associated with who receives the treatment of interest and who does not.

This analysis highlights the risks of using propensity scores where they do not adequately adjust for confounding by indication. Propensity scores, like all observational methods, need to be developed and used with caution, and results from propensity analyses need to be interpreted with caution. A useful starting point, as was taken in this example, is to start with the replication of a known treatment effect, and bridge from there to consider further unanswered questions. Considering the interaction between the propensity score and treatment outcome may also be a useful step to aid understanding of the score.

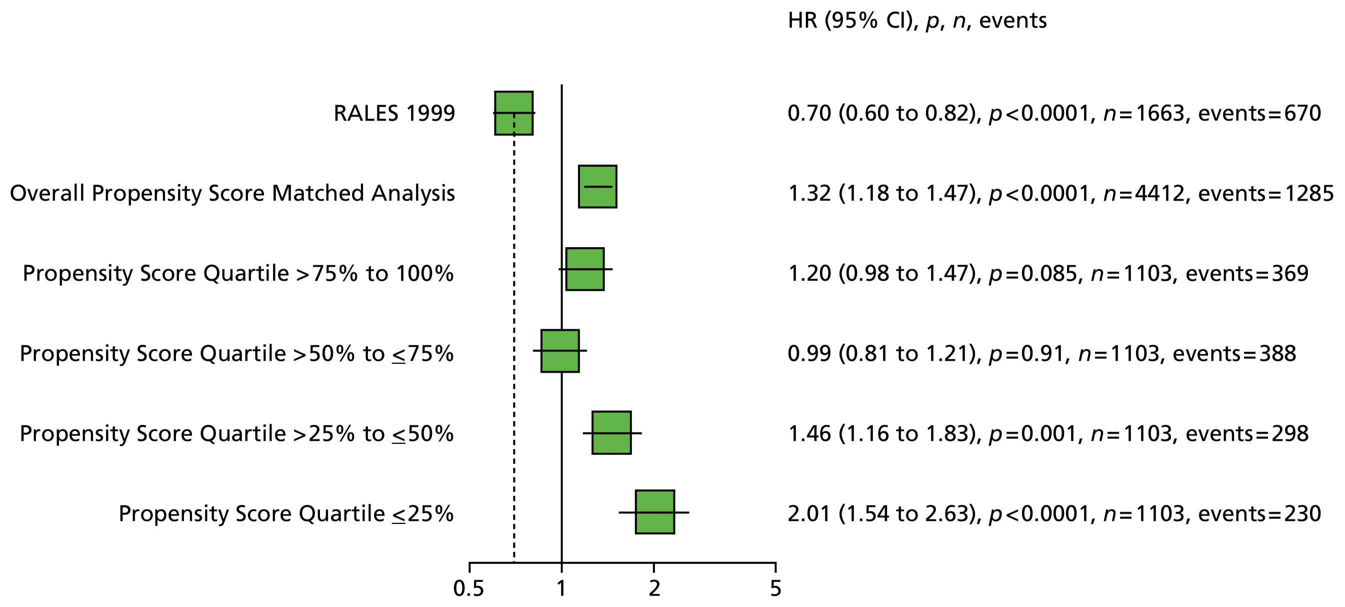

FIGURE 3.4 Results of the propensity score analysis (overall and in quartiles) compared with the results of the RALES trial. HR, hazard ratio. Adapted by permission from BMJ Publishing Group Limited. Making interferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research, Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I, vol. 347, p. f6409, © 2013.[19]

More recent studies have suggested that alternative approaches to generating a matched control group may perform better. An approach that may hold particular promise is 'genetic matching', which does not assume that there is a propensity score which is correctly specified but rather uses a multivariate matching (genetic) algorithm to produce a matched control group to maximise the balance of observed covariates between the treatment and control groups.[23]

Moving forward with quantitative methods for evaluating complex interventions

The growth in the availability of administrative, survey and other forms of 'observational' data offers huge potential for improved and more comprehensive evaluation of health and care interventions. Experimental trials, although allowing the use of 'gold-standard' methods such as RCTs, are often very expensive and time-consuming. Moreover, as outlined in this paper, the findings of RCTs may be limited with regard to generalisation: that is, feature reduced externality validity. Observational quantitative approaches can offer alternative methods of evaluation, particularly for complex interventions, by helping to both address issues of generalisability of RCT results and provide direct estimates of treatment effects. Methods for observational or non-randomisation studies have been advancing substantially in the literature, with many innovations now available to better address core challenges such as selection bias, and therefore to better improve internal validity Other essays in this volume (particularly *Essays 6* and *7*) describe other features of optimal study designs to evaluate complex features of health and care systems.

Observational studies will always be subject to treatment selection bias to some extent, and so have reduced internal validity compared with RCTs. Nonetheless, methods such as matching, synthetic control and IVs may mitigate the problem, if not completely eradicate any biases, especially when there are unknown confounding factors. Observational studies, therefore, can be used in place of RCTs to estimate treatment effects if careful consideration is given to the analysis and it is accepted that some bias will remain.[24,25] Observational methods are also being used alongside RCTs to directly tackle the issue of representativeness of experimental studies.

The growing availability of routine data allows the use of observational methods. Although the quality, availability and completeness of routine data could always be improved to allow its easier use for the assessment of complex interventions, routine data can be used at relatively low cost and often have large scale and breadth (if not depth). Arguably, routine data sets are an underutilised resource, and greater investment to improve routine data sets and make them more research-friendly would increase their usage. Of course, routine data have their limitations: primary outcomes of interest may not be recorded, and there may be issues with the completeness and compatibility of data sets. In general, though, they have high external validity as they draw on data about everyday operation, and they can accommodate a high degree of subgroup analysis to explore why and when an intervention works. More pragmatically, observational studies based on routine data are generally low cost and have high feasibility.

Quantitative methods for estimating the causal effects of complex intervention inevitably make strong assumptions which must be critically examined. Future studies should report effectiveness estimates according to approaches that make different but plausible underlying assumptions. Where new methods are being utilised, it would be useful to compare the results with those from standard or appropriate alternative methodologies, and implications of the chosen analysis should also be explored through sensitivity analyses.

There is a growing need to demonstrate effectiveness and cost-effectiveness of complex interventions. Although just scratching the surface, this essay has shown that evaluators have a range of quantitative methods available, which include those in both the experimental and the observational toolboxes. However, it is perhaps the use of a combination of these approaches that might be most suited to evaluating complex interventions.

Acknowledgements

The authors would like to thank the editors; co-authors of some of the paper cited in this essay, including Noemi Kreif, Matt Sutton, Erin Hartmann and Adam Steventon; and participants at the Evaluation London 2015 workshop. We would also like to thank Jane Dennett for her help in copy-editing and preparing this manuscript. Any errors and omissions are the responsibility of the authors.

Contributions of authors

Clare Gillies (Medical Statistician) synthesised the evidence and literature, and produced the first draft of the essay.

Nick Freemantle (Epidemiologist and Biostatistician) provided expert input (particularly on the validity of RCTs and matching) and helped to edit the essay.

Richard Grieve (Health Economics Methodologist) provided expert input (particularly on observational methods) and helped to edit the essay.

Jasjeet Sekhon (Political Scientist and Statistician) provided expert input (particularly on RCT design and observational methods) and helped to edit the essay.

Julien Forder (Economist) edited the essay, added further material, consolidated the various inputs and produced the final draft.

References

1. Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. *BMJ* 2008;**336**:1281–3. http://dx.doi.org/10.1136/bmj.39569.510521.AD
2. Hartman E, Grieve R, Ramsahai R, Sekhon JS. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *J Roy Stat Soc A Sta* 2015;**178**:757–78. http://dx.doi.org/10.1111/rssa.12094
3. Bower P, Cartwright M, Hirani SP, Barlow J, Hendy J, Knapp M, *et al.* A comprehensive evaluation of the impact of telemonitoring in patients with long-term conditions and social care needs: protocol for the Whole Systems Demonstrator cluster randomised trial. *BMC Health Serv Res* 2011;**11**:184. http://dx.doi.org/10.1186/1472-6963-11-184
4. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, *et al.* Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. *BMJ* 2012;**344**:e3874. http://dx.doi.org/10.1136/bmj.e3874
5. Steventon A, Grieve R, Bardsley M. An approach to assess generalizability in comparative effectiveness research: a case study of the whole systems demonstrator cluster randomized trial comparing telehealth with usual care for patients with chronic health conditions. *Med Decis Mak* 2015;**35**:1023–36. http://dx.doi.org/10.1177/0272989X15585131
6. Jones K, Forder J, Caiels J, Welch E, Glendinning C, Windle K. Personalization in the health care system: do personal health budgets have an impact on outcomes and cost? *J Health Serv Res Policy* 2013;**18**(Suppl. 2):59–67. http://dx.doi.org/10.1177/1355819613503152
7. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J Am Med Assoc* 2010;**105**:493–505. http://dx.doi.org/10.1198/jasa.2009.ap08746

8. Kreif N, Grieve HD, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units [published online ahead of print 7 October 2015]. *Health Econ* 2015.

9. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Ann Rev Public Health* 1998;**19**:17–34. http://dx.doi.org/10.1146/annurev.publhealth.19.1.17

10. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on ami survival using propensity score and instrumental variable methods. *J Am Med Assoc* 2007;**297**:278–85. http://dx.doi.org/10.1001/jama.297.3.278

11. Jones AM, Rice N. Econometric Evaluation of Health Policies. In Glied S, Smith PC, editors. *The Oxford Handbook of Health Economics*. Oxford: Oxford University Press; 2011. http://dx.doi.org/10.1093/oxfordhb/9780199238828.013.0037

12. Forder J, Malley J, Towers AM, Netten A. Using cost-effectiveness estimates from survey data to guide commissioning: an application to home care. *Health Econ* 2014;**23**:979–92. http://dx.doi.org/10.1002/hec.2973

13. Netten A, Burge P, Malley J, Potoglou D, Towers A. Outcomes of social care for adults: developing a preference weighted measure. *Health Technol Assess* 2012;**16**(16). http://dx.doi.org/10.3310/hta16160

14. Claxton K, Martin S, Soares M, Rice N, Spackman E, Hinde S, *et al.* Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. *Health Technol Assess* 2015;**19**(14). http://dx.doi.org/10.3310/hta19140

15. Forder J. Long-term care and hospital utilisation by older people: an analysis of substitution rates. *Health Econ* 2009;**18**:1322–38. http://dx.doi.org/10.1002/hec.1438

16. Gaughan J, Gravelle H, Siciliani L. Testing the bed-blocking hypothesis: does nursing and care home supply reduce delayed hospital discharges? *Health Econ* 2015;**24**:32–44. http://dx.doi.org/10.1002/hec.3150

17. Angrist JD, Pischke J-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press; 2009.

18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55. http://dx.doi.org/10.1093/biomet/70.1.41

19. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;**347**:f6904. http://dx.doi.org/10.1136/bmj.f6409

20. Pitt B, Zannad F, Remme WJ, Cody R, Castaigne A, Perez A, *et al.* The effect of spironolactone on morbidity and mortality in patients with severe heart failure. *N Engl J Med* 1999;**341**:709–17. http://dx.doi.org/10.1056/NEJM199909023411001

21. Pitt B, Remme W, Zannad F, Neaton J, Martinez F, Roniker B, *et al.* Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *N Engl J Med* 2003;**348**:1309–21. http://dx.doi.org/10.1056/NEJMoa030207

22. Zannad F, McMurray JJV, Krum H, van Veldhuisen DJ, Swedberg K, Shi H, *et al.* Eplerenone in patients with systolic heart failure and mild symptoms. *N Engl J Med* 2011;**364**:11–21. http://dx.doi.org/10.1056/NEJMoa1009492

23. Diamond A, Sekhon JS. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev Econ Stat* 2013;**95**:932–45. http://dx.doi.org/10.1162/REST_a_00318

24. Freemantle N, Richardson M, Wood J, Ray D, Khosla S, Shahian D, *et al.* Weekend hospitalization and additional risk of death: an analysis of inpatient data. *J Roy Soc Med* 2012;**105**:74–84. http://dx.doi.org/10.1258/jrsm.2012.120009

25. Lester W, Freemantle N, Begaj I, Ray D, Wood J, Pagano D. Fatal venous thromboembolism associated with hospital admission: a cohort study to assess the impact of a national risk assessment target. *Heart* 2013;**99**:1734–9. http://dx.doi.org/10.1136/heartjnl-2013-304479

Essay 4 Patient-reported outcome measures and the evaluation of services

Elizabeth Gibbons,[1] Nick Black,[2] Lesley Fallowfield,[3] Robin Newhouse[4] and Ray Fitzpatrick[1]

[1]Health Services Research Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, UK
[3]Sussex Health Outcomes Research and Education in Cancer (SHORE-C), University of Sussex, Brighton, UK
[4]Indiana University School of Nursing, Indianapolis, IN, USA

Declared competing interests of authors: Robin Newhouse reports personal fees from the Patient Centered Outcomes Research Institute and being chairperson and on the methodology committee, being a board member for AcademyHealth, and grants from Agency for Health Care Research and Quality outside the submitted work.

Published May 2016
DOI: 10.3310/hsdr04160-55

This essay should be referenced as follows:

Gibbons E, Black N, Fallowfield L, Newhouse R, Fitzpatrick R. Patient-reported outcome measures and the evaluation of services. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 55–68.

List of boxes

List of abbreviations

CAT	computer adaptive testing
PCORI	Patient-Centered Outcomes Research Institute
PROM	patient-reported outcome measure
PROMIS	Patient Reported Outcomes Measurement Information System
RCT	randomised controlled trial

Abstract

A growing consensus has emerged about patient-reported outcome measures (PROMs) as an important tool in the evaluation of services. There is little dispute regarding the range of relevant dimensions of health status, the types of measures available to capture what matters to patients, their required measurement properties and reporting standards when PROMs are used in studies. Their use on a large scale in the NHS national PROMs programme has produced valuable lessons about elective surgery but also about PROMs as tools.

There are outstanding issues if PROMs are to be further applied in the evaluation of services. Patients need to be actively engaged. Health professionals need to see the merits of this approach to patient care and outcome assessment. PROMs need to be integrated into health records. Some recent initiatives have attempted to address these three needs, and further development and testing of such initiatives is needed.

Scientific summary

A growing consensus has emerged about patient-reported outcome measures (PROMs) as important tools in the evaluation of services. There is little dispute regarding the range of relevant dimensions of health status, the types of measures available to capture what matters to patients, their required measurement properties and reporting standards when PROMs are used in studies. Their use on a large scale in the NHS national PROMs programme has produced valuable lessons about elective surgery but also about PROMs as tools.

Respondent engagement is fundamental to ensure adequate response rates to be able use outcomes for the evaluation of services. Evidence exists that patients with poorer health status at baseline impact on response rates which may produce misleading evidence of the effectiveness of services. Different mechanisms for the delivery and capture of PROMs are being developed such as electronic data collection and use of other media. This could complement existing paper-based methods and has the potential to reduce respondent burden of completion.

Specific populations pose particular challenges in engagement such as those with developmental limitations (children) and those with cognitive impairment such as patients with dementia or learning difficulties. In addition, the collection of baseline measurement in patients experiencing acute or emergency health problems poses difficulties.

Missing or incomplete data present analytical problems and, despite a range of statistical procedures used, performance of services could be influenced by methods adopted.

Evidence suggesting there is diversity in levels of staff engagement from enthusiasm to scepticism. Difficulties are apparent in relation to understanding and interpretation of the data, as well as the challenges of incorporating data into existing systems and processes of care. This results in little impact of PROMs data used for service improvement.

Further evolution in PROMs is still needed so that they focus as much on practical, feasible and clinically actionable content as on the psychometrically validated content emphasised to date. Development and testing of feasible software, platforms and electronic health records needs to be developed to support PROMs as a routine feature of both clinical care and evaluative research. Training of staff to strengthen understanding of PROMs and interpreting will be essential. Most outstanding of all will be the need to establish benefits to patients. There is much speculation about PROMs supporting and informing patients' choice of health-care options, health and consulting behaviour and self-management, but there are very few studies to demonstrate such benefits.

Introduction

Health services are under constant pressure to be more patient centred and positively valued by patients and the public. Evaluative research reflects that pressure by increasingly assessing outcomes as perceived by patients. Patient-reported outcome measures (PROMs) are a diverse array of questionnaires and related techniques intended to obtain patients' own views of their health status and benefits experienced from receiving health services. Because of their focus on accurately capturing the patient's perspective on outcomes, they have been seen for some time as having enormous and distinctive potential to transform the assessment of the performance of services.[1] A recent overview described measures of health status (effectively PROMs) as '. . . the most important scientific development in the last 50 years in the field of health services research'.[2] This essay provides an overview of the progress that has been made in the measurement of patients' views of health status and outcomes, and identifies some of the key challenges to be faced in their further use to evaluate services.

Applications of patient-reported outcome measures

Patient-reported outcome measures have been considered as solutions to a diverse range of problems. First developed over 40 years ago, they were early on considered as invaluable outcome measures in health services research, for example in the widely cited randomised controlled trial (RCT) of alternative mechanisms of reimbursement, the RAND Health Insurance Experiment.[3] Recent examples of the use of PROMs as outcomes in large multicentre health services research include their use to judge the benefits of personally controlled health budgets,[4] and as outcomes of telehealth services.[5] It is in the adjacent field of clinical trials that PROMs have been most widely applied, often as secondary outcomes to weigh against traditional indicators such as mortality but occasionally as the primary end point. A recent survey of nearly 100,000 clinical trials published between 2007 and 2013 found that a PROM was used in 27% of trials.[6] A specific subset of such trials includes cost–utility analysis, estimating utility or net value of a treatment by means of a particular type of PROM described in *Types of patient-reported outcome measure* as utility measure.

Another use of the PROM is to provide alternative expressions of health need in the population. A classic example is the use of a measure, a version of the Sickness Impact Profile, to assess the epidemiology of disability in London.[7] A related application is their use as prognostic variables and impressive evidence exists for this use in areas such as cancer.[8]

Patient-reported outcome measures are beginning to be used as instruments to contribute to local quality improvement developments; for example, Nilsson *et al.*[9] recently reviewed the use of PROMs in Swedish disease registers and found PROMs commonly used this way. There is growing interest in whether or not they have a role in improving individual patient care, for example by strengthening shared decision-making or improving the detection and management of patients' problems. Trials to date have had mixed results.[10,11] As will be argued later, their acceptance in routine patient care may prove to be crucial to furthering their use in evaluative research.

The example of cancer provides compelling evidence of why validated PROMs are needed to evaluate the performance of services. Generally, health professionals show poor understanding of how cancer affects quality of life, both in regular care and in evaluative research. In clinical trials for ovarian cancer, patients reported more symptoms and reduced quality of life compared with clinician-completed toxicity scales.[12] Vera-Badillo *et al.*[13] reviewed a series of breast cancer trials and found both bias and under-reporting of toxicities. Di Maio *et al.*[14] compared the reports by patients and physicians of six specific toxicities reported by patients and physicians (anorexia, nausea, vomiting, constipation, diarrhoea and hair loss) in three RCTS for cancer. Physicians consistently under-recognised and under-reported toxicities. Such results mean not only that the benefit-to-harm ratio of treatments is often inaccurate but also that the higher burdens on patients experiencing these symptoms can lead to non-adherence or dose reductions limiting the expected clinical benefits.

Some recent developments have raised the visibility of PROMs as a distinctive lens through which to assess the quality of health services. In England, a national PROMS programme was introduced in 2009 mandating the use of PROMs for NHS patients in relation to selected elective surgical procedures. This is described further in *A case study: patient-reported outcome measures and NHS elective surgery* because of its very size, ambition and potential lessons.

In the USA, major impetus for PROMs has stemmed from the establishment with major funding by Congress of the Patient-Centered Outcomes Research Institute (PCORI) designed to answer key uncertainties in health-care delivery by the conduct of patient-centred comparative effectiveness research. A key feature of the work of PCORI is that patients and patients' values should be central to every stage of health research.[15] Not surprisingly, the application and improved methods for patient-reported outcomes are therefore a priority (www.pcori.org/events/2013/patient-reported-outcomes-pro-infrastructure-workshop).

Developing consensus

There is substantial and growing consensus about PROMs. This essay focuses on four key areas of agreement: relevant dimensions, required measurement properties, types of measure and reporting standards.

Relevant dimensions

A key feature of PROMs is that they assess health in the broadest sense. Individual instruments vary in content but the domains identified in *Box 4.1*[16] cover most aspects of health status addressed by PROMs. In their most common form, measures of dimensions of health status, such as are shown in *Box 4.1*, are assessed before and after an intervention, and, strictly, it is the change over time that renders the expression of an outcome that, in principle, may be attributed to an intervention.

BOX 4.1 Dimensions of health in PROMs

Physical function

Mobility, dexterity, range of movement, physical activity

Activities of daily living: ability to eat, wash, dress

Symptoms

Pain, nausea, appetite, energy, vitality, fatigue, sleep and rest

Global judgements of health

Psychological well-being

Psychological illness: anxiety, depression

Coping, positive well-being and adjustment, sense of control, self-esteem

Social well-being

Family and intimate relations, social contact, integration, and social opportunities, leisure activities, sexual activity and satisfaction

Cognitive functioning

Cognition, alertness, concentration, memory, confusion, ability to communicate

Role activities

Employment, household management, financial concerns

Personal constructs

Satisfaction with bodily appearance, stigma and stigmatising conditions, life satisfaction, spirituality

Satisfaction with care

Reproduced with permission from Fitzpatrick *et al.*[16] Contains information licensed under the Non-Commercial Government Licence v1.0' from this link: http://www.journalslibrary.nihr.ac.uk/rights_and_permissions.

Measurement properties

There is broad agreement about the measurement properties required of a PROM and techniques used to assess such properties. Reliability is the extent to which a PROM is free from measurement error and reproducible. Face and content validity concern basic judgements about whether or not items of a questionnaire appear to address the experiences relevant to respondents. In practice, this is commonly judged from evidence of the extent to which patients have fully contributed to the derivation and approval of content of an instrument. Construct validity is evaluated to explore the hypothetical constructs it purports to measure and ensure that there is a statistical relationship between measures of similar constructs. The measure should avoid ceiling or floor effects demonstrating precision of measurement. Consideration of the ability of the measure to detect change and ensure responsiveness is particularly important for PROMs, most usefully tested against the respondent's own retrospective judgement of change. When used to determine the 'value' of treatments from an international perspective, cultural sensitivity and appropriate forwards and backwards field-testing of translations are also vital to demonstrate.

These properties are familiar from other fields of measurement in science. They are tested by a range of methods familiar from classical psychometrics as well as more recently applied techniques such as Rasch Analysis and Item Response Theory.[17]

However, as PROMs are almost invariably used with patients receiving health care for real presenting problems, it is as important to consider the acceptability of PROMs along with their technical measurement properties in order to minimise burden. Minimal burden should also strengthen response rates and hence the generalisability of results.

Types of patient-reported outcome measure

There is broad agreement about the range of types of PROM available and their respective merits.

Generic PROMs are more commonly multidimensional and applicable across a range of conditions, and enable comparison within and between different health conditions, populations and interventions. The SF-36 (Short Form Questionnaire-36 items) is probably the most widely used such measure.

Preference-based measures, such as the European Quality of Life-5 Dimensions (EQ-5D) (www.euroqol.org/), provide a descriptive profile and index of a person's health status in the form of a utility derived from public preferences for different health states and weighted accordingly. The distinctive role of such measures is therefore used in economic evaluations such as cost–utility analyses. They are a specific subtype of generic measure.

Disease- and or procedure-specific PROMs aim to represent particular aspects of health status relevant, for example, to patients living with diabetes or those undergoing a specific procedure, such as joint replacement. The distinctive merit of such measures is to be maximally sensitive to the particular challenges posed by a given disease and, therefore, maximally sensitive to benefits of an intervention for the condition.

Individualised measures of health status elicit patients' personal goals and concerns regarding their health, thereby avoiding the standardised format of the other measures just described.

A common strategy in the evaluation of a specific intervention for a limited range of conditions is to recommend that an evaluation include a generic and a disease-specific measure to address the widest spectrum of anticipated and unanticipated outcomes.

Reporting standards

The field has evolved to the extent that a range of guidelines has now been developed relevant to many aspects of PROMs and their use. Thus, several influential guidelines have emerged regarding appropriate methods and standards for the development of a PROM.[18,19] Other guidance focuses on how PROMs should be reported.[20] Guidance also exists on how to carry out systematic reviews of the measurement properties of PROMs.[21] Although invaluable for the development of the field, implicit in much of such guidance is a focus on clinical trials for single treatments. Much routine assessment of services needs to be based on more complex interventions delivered to heterogeneous populations, for which such guidance may be less helpful.

Strengthening use in health services research

A number of developments have occurred that may enhance the role of PROMs in health services research. Thus, studies have begun to show that shorter versions of PROMs may have equivalent measurement properties, a development that may significantly decrease burden and increase response rates.[22] Caution must be used in methods of shortening to avoid losing key content.[23] Juniper *et al.*[24] proposed taking account of the frequency with which problems are reported in different items, combined with patients' ratings of their importance, to select key items.

There may be circumstances when, for example, disease-specific PROMs have been used in the valuation of an intervention but there has been no preference-based measure, making cost–utility analyses difficult. A range of techniques have now been developed that map utility values across to non-preference-based measures; the advantage is that not only can utility values be calculated but potential additional precision of the disease-specific measure is retained.[25]

Methods for interpreting scores from PROMs have advanced. The most important scores generated by PROMs are usually the changes over time from before to after an intervention, but numerical change scores of non-intuitive scales may be difficult to interpret. Traditional solutions focused purely on statistical solutions to establish minimally important differences. However, an alternative is to identify the smallest change that would be appreciated especially by patients, identified by relating change scores to patients' retrospective judgements of how they value change. This approach, referred to as 'anchor-based', has the advantage of determining the meaning of PROMs from the patient's perspective, which can still be cross-checked with more statistical approaches.[26]

The content of PROMs relevant to evaluative research is also being expanded to include related valued goals and concerns of the patient. Although some may challenge the extension of PROMs beyond health status, a range of related constructs may be just as important to the patient and closely related in potential causal chains between intervention and final health outcome, for example self-efficacy and patient activation. Additionally, the boundary between PROMs and patient experience may also need to be relaxed; when, for example, patients are capable of giving meaningful answers to validated questionnaires retrospectively judging the outcomes and benefits of treatments.[27–29] An ongoing problem involves the phenomenon of response-shift bias, which occurs when a patient with a deteriorating condition nevertheless reports an unchanged quality of life owing to successful adaptation and a shift in the thresholds they themselves use to describe the severity or impact of a problem.

A case study: patient-reported outcome measures and NHS elective surgery

Since 2009, the providers of care for NHS patients have been required to collect data before and 3 or 6 months after surgery with a condition-specific and a generic PROM for four elective procedures: hip replacements, knee replacements, groin hernia repair and varicose veins surgery. Importantly, the information generated is publicly available (www.hscic.gov.uk/proms) and guidance is provided to help the public navigate and interpret data for individual hospital trusts. The database has provided a uniquely rich source for evaluative research using the PROM.

Some overall patterns are striking; for example, the evidence for the effectiveness of joint replacement (especially in terms of condition-specific measures) and the relative lack of evidence of provider trusts being consistently poor outlier performers in relation to appropriately adjusted PROMs are reassuring for the NHS. There was somewhat more evidence of individual surgeons performing markedly poorer, according to PROMs, compared with the much lower variation according to 90-day post-surgical mortality.[30] The national PROMs programme provide results that challenge conventional thinking. There is no support for the view that either individual surgeons or whole trusts with larger volumes of surgery have better results.[31]

The programme has provided important methodological insights, showing how response rates and choice of PROM may influence the comparative performance of trusts. As discussed below, these results illustrate the outstanding challenges for future use of PROMs.

Major challenges

Respondent engagement

Patient-reported outcome measures are unusual compared with almost all other health indicators, in that they require active input from the patient or respondent. For example, in the NHS national PROMs programme, whereas pre-operative recruitment rates to PROMs for patients receiving joint replacement surgery have been 68%, the rate is markedly lower, 41%, in patients receiving varicose vein repair surgery, although this is largely attributable to eligible patients not being invited.[32] In a sample of patients recruited to complete PROMs for any of six long-term conditions recruited via general practices, the recruitment rate was lower again, at 38%.[33] There is some evidence that response rates have deteriorated over time; Hazell *et al.*[34] found marked deteriorations in response rate to an identical survey to patients about their asthma, from 71% in 1993 to 47% in 2004.

Clearly, a number of factors may influence the response rate. In the examples just cited, patients were more likely to respond to surveys about receiving a specific surgical intervention than to primary care surveys about long-term conditions where there was no link to receiving treatment. The concern is that the lower the response rate, the harder it is to use the outcomes obtained to evaluate services.

However, the greater concern is if there is evidence of response bias: if important characteristics of either patients or services are associated with differential response rate. There is evidence that this may occur for PROMs. In the national PROMs programme, Hutchings *et al.*[35] found that the poorer the health status in terms of comorbidities prior to surgery, the poorer the response rate to PROMs questionnaires mailed out after surgery. Similarly, in the primary care study of PROMs for long-term conditions referred to above, when patients were asked to return a follow-up PROM 1 year after the baseline, the response rate was lower in those with poorer baseline health status.[36] It is clear that such biases may produce misleading evidence of the effectiveness of services.

There are more subtle forms of problem with PROMs if questionnaires are returned incomplete, for example if more sensitive or personal questions are not completed or more difficult items are omitted. Usually the development phase of PROMs reduces such risks. There is a range of statistical techniques for addressing missing or incomplete data from longitudinal data sets such as PROMs, most commonly the use of imputation methods derived from the information about respondents that has been obtained. Analysing missing data from the national PROMs programme, Gomes *et al.*[37] found that inferences about the performance of services could be influenced by the assumptions and methods made to make imputations to address missing data.

To address the problem of non-response, most effort has gone into the design of PROMs and mechanisms of delivery (e.g. the use of the internet and other modern media), partly because these are practical solutions that can be implemented. The scope for innovative technology to enhance respondent engagement is significant, although, currently, for the majority of PROMs, evaluative research still relies on traditional survey methods, particularly the mailed questionnaire. Although beyond the scope of this essay, recent innovations in the core format of PROMs may also strengthen their acceptability and power to engage the respondent. The innovation is the use of computer adaptive testing (CAT) to tailor questionnaire items to the individual, with responses to initially administered items determining the choice of subsequent items. The total number of items required to be completed is significantly reduced. This has been the subject of a major National Institutes of Health-funded initiative in the USA, Patient Reported Outcome Measures System (PROMIS) (www.nihpromis.org/), to produce questionnaire items that can populate CAT systems. Although an exciting initiative in the science of PROMs, particularly in terms of their use in assessing individuals' health, there is little evidence to date of their use in evaluative research.

The other major concern occurs where there are major difficulties in engaging respondents, whether because of major physical, cognitive or developmental limitations or because of social exclusion. One example is the involvement of children. PROMs exist that are designed to be relevant to the perceptions of children.[38] Additional care may be required, in terms of appropriate interviewing, determining intellectual capacity, not relying solely on chronological age, and the role of observation and proxy informants.[39]

Dementia poses related challenges. There are dementia-specific PROMs, which appear to work satisfactorily for patients with mild to moderate dementia.[40] However, for those with more severe levels of dementia, assessments by a proxy carer are the more plausible option. Nevertheless, proxy ratings should not be treated simplistically or uncritically because their rates may, in turn, be influenced by burden of care and carer burnout, and appropriate adjustments are, therefore, required.[41]

In some areas of health care it is impossible to obtain a pretreatment or baseline assessment for obvious reasons; patients who experience sudden health events such as a stroke or hip fracture will not have any reason to have completed such an assessment. A number of approaches may be adopted for which the patient's perspective is needed. One approach is simply to give up on pre-event assessment and assess progress and possible impact of interventions after the sudden event and as soon as it is feasible to engage the patient.[42] Alternative strategies can include inviting the patient retrospectively to assess their pre-event health or to use appropriately matched population-based data. Neither of these two strategies is straightforward.

There is only limited evidence of the impact of social exclusion in relation to PROMs. The problem is, however, well highlighted in a study by Jahagirdar *et al.*,[43] who found that excluded groups, such as those with learning difficulties or low literacy, were less likely to be involved in the development of PROMs for chronic obstructive pulmonary disease. The commonest pragmatic solution to social exclusion more generally is to weight evidence for under-represented groups in surveys. A recent review of the sparse evidence argues that multiple and flexible approaches are needed properly to ensure that socially excluded groups' views are not overlooked[44] (www.pssru.ac.uk/archive/pdf/4390.pdf). The importance of social inequalities and social exclusion in health services research is further discussed in *Essay 5* in this volume.

Health professional engagement

If the role of PROMs is to be expanded, it is likely to require greater engagement of clinicians so that collection of PROMs becomes more a part of routine care. Reference has already been made to evidence from trials introducing PROMs into individual patient care and showing mixed evidence of impact on management decisions, patient experience of care and ultimate health outcomes. This variable impact may, in turn, be attributable to health professionals' attitudes, beliefs and experiences regarding PROMs; these have been the subject of some research. A recent study of surgeons' views of PROMs found considerable diversity with a range from enthusiast to sceptic.[45] In this and other studies, a number of concerns have been expressed. Cognitive problems include the view that PROMs data are difficult to interpret and relate to management decisions.[46,47] Other concerns focus on logistics, time constraints and difficulties of incorporating PROMs into clinical routines.[46–48] These studies generally stress clinicians' expressed need for greater training to incorporate PROMs into practice.[49] At worst, some studies suggest concerns that PROMs actually may cause harm if evidence from them is misunderstood or misused by third-party audiences such as managers, commissioners or politicians.[50,51]

Patient-reported outcome measures and routine health-care revisited

A recent survey obtained the views of relevant experts from the USA, England and the Netherlands about prospects for future use and impact of PROMs.[52] There was a clear consensus that for PROMs to have their fullest impact in assessing the performance of health services, there needed to be greater integration of information systems for routine patient care and for system performance measurement. Currently in all three countries, information systems for the two different functions are effectively independent. The experts identified a number of barriers that would need to be overcome to integrate these two worlds of activity, including lack of trust from participants and insufficient belief in the value of PROMs in patient care.

A number of reports are beginning to appear broadly supportive of the conclusions of Van der Wees *et al.*[52] and also express optimism that the integration of information from PROMS for patient care and system performance can be achieved. In the USA, Wu *et al.*[53] describe a number of health record systems already available and in use to support the two functions. They identify three developments favourable to this integration: the positive trend towards patient-centeredness and electronic applications for PROMs, the growth of electronic health records and trend towards comparative effectiveness research that is patient oriented. More recently, Jensen *et al.*[48] reported a number of encouraging case studies in the USA in which PROMs served both patient care and system evaluation. They effectively identify the same three generally supportive developments as Wu *et al.*[53]

Warrington *et al.*[54] provide a positive account of a NHS setting in which PROMs have been successfully implemented to provide long-term follow-up of cancer survivors. They acknowledge that in a systematic review of their field of cancer care, Nama *et al.*[55] could find no high-quality evaluations to demonstrate clearly the benefits of such systems. This lack of evidence remains a major challenge.

Conclusion

Patient-reported outcome measures have an important role in evaluating services in terms of outcomes that matter to patients. They will continue to play a role as primary or secondary end points in bespoke research trials and evaluative studies in which the patient is recruited to participate as research subject. In addition, there is likely to be a larger role of patients' routinely contributing information about their health and outcomes of their care as health-care systems adapt to take advantage of developments in informatics and PROMs. In anticipation of such developments, it has been argued that further evolution in PROMs is still needed so that they focus as much on practical, feasible and clinically actionable content as on the psychometrically validated content emphasised to date.[56] Further development and testing is needed of feasible software, platforms and electronic health records to support PROMs as a routine feature of both clinical care and evaluative research. Major issues of trust, confidentiality and sustainability remain to be addressed. As already emphasised, training to strengthen understanding will be essential. Evaluative studies will be needed to test the overall viability of electronic health-based health records that include PROMs. Most outstanding of all will be the need to establish benefits to patients. There is much speculation of PROMs supporting and informing patients' choice of health-care options, health and consulting behaviour and self-management, but there are very few studies to demonstrate such benefits.

Acknowledgements

Contributions of authors

Elizabeth Gibbons (Senior Research Scientist, Health Services Research) and **Ray Fitzpatrick** (Professor, Public Health) wrote the first draft of the essay.

Nick Black (Professor, Health Services Research), **Lesley Fallowfield** (Professor, Psycho-oncology) and **Robin Newhouse** (Professor, Nursing) commented on and contributed to the draft.

References

1. Elwood P. Shattuck lecture – outcomes management. A technology of patient experience. *N Engl J Med* 1988;**318**:1549–56.
2. Brook R, Vaiana M. Using the knowledge base of health services research to redefine health care systems. *J Gen Intern Med* 2015;**30**:1547–56. http://dx.doi.org/10.1007/s11606-015-3298-2
3. Newhouse J. A summary of the RAND Health Insurance study. *Ann N Y Acad Sci* 1982;**387**:111–14. http://dx.doi.org/10.1111/j.1749-6632.1982.tb17166.x
4. Jones K, Forder J, Caiels J, Welch E, Glendinning C, Windle K. Personalization in the health care system: do personal health budgets have an impact on outcomes and cost? *J Health Serv Res Policy* 2013;**18**(Suppl. 2):59–67. http://dx.doi.org/10.1177/1355819613503152
5. Cartwright M, Hirani SP, Rixon L, Beynon M, Doll H, Bower P, *et al.* Whole Systems Demonstrator Evaluation Team. Effect of telehealth on quality of life and psychological outcomes over 12 months (Whole Systems Demonstrator telehealth questionnaire study): nested study of patient reported outcomes in a pragmatic, cluster randomised controlled trial. *BMJ* 2013;**346**:f653. http://dx.doi.org/10.1136/bmj.f653
6. Vodicka E, Kimb K, Devinea E, Gnanasakthy A, Scoggins J, Patrick D. Inclusion of patient-reported outcome measures in registered clinical trials: evidence from ClinicalTrials.gov (2007–2013). *Contemp Clin Trials* 2015;**43**:1–9. http://dx.doi.org/10.1016/j.cct.2015.04.004
7. Patrick D, Peach H. *Disablement in the Community*. Oxford: Oxford University Press; 1989.

8. Gotay CC, Kawamoto CT, Bottomley A, Efficace F. The prognostic significance of patient-reported outcomes in cancer clinical trials. *J Clin Oncol* 2008;**26**:1355–63. http://dx.doi.org/10.1200/JCO.2007.13.3439

9. Nilsson E, Orwelius L, Kristenson M. Patient-reported outcomes in the Swedish National Quality Registers. *J Int Med* 2015;**279**:141–53. http://dx.doi.org/10.1111/joim.12409

10. Marshall S, Haywood K, Fitzpatrick R. Impact of patient-reported outcome measures on routine practice: a structured review. *J Eval Clin Pract* 2006;**12**:559–68. http://dx.doi.org/10.1111/j.1365-2753.2006.00650.x

11. Valderas JM, Kotzeva A, Espallargues M, Guyatt G, Ferrans CE, Halyard MY, *et al.* The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Qual Life Res* 2008;**17**:179–93. http://dx.doi.org/10.1007/s11136-007-9295-0

12. Greimel ER, Bjelic-Radisic V, Pfisterer J, Hilpert F, Daghofer F, Pujade-Lauraine E, *et al.* Toxicity and quality of life outcomes in ovarian cancer patients participating in randomized controlled trials. *Support Care Cancer* 2011;**19**:1421–7. http://dx.doi.org/10.1007/s00520-010-0969-8

13. Vera-Badillo FE, Shapiro R, Ocana A, Amir E, Tannock IF. Bias in reporting of end points of efficacy and toxicity in randomized, clinical trials for women with breast cancer. *Ann Oncol* 2013;**24**:1238–44. http://dx.doi.org/10.1093/annonc/mds636

14. Di Maio M, Gallo C, Leighl NB, Piccirillo MC, Daniele G, Nuzzo F, *et al.* Symptomatic toxicities experienced during anticancer treatment: agreement between patient and physician reporting in three randomized trials. *J Clin Oncol* 2015;**33**:910–15. http://dx.doi.org/10.1200/JCO.2014.57.9334

15. Newhouse R, Barksdale DJ, Miller JA. The Patient-Centered Outcomes Research Institute: research done differently. *Nurs Res* 2015;**64**:72–7. http://dx.doi.org/10.1097/NNR.0000000000000070

16. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials: a review. *Health Technol Assess* 1998;**2**(14).

17. Streiner D, Norman G. *Health Measurement Scales*. Oxford: Oxford University Press; 1995.

18. Food and Drug Administration Department of Health and Human Sciences. *Guidance to Industry Patient Reported Outcome Measures Use in Medical Product Development to Support Labelling Claims*. Silver Spring, MD: Department of Health and Human Sciences; 2009.

19. Reeve B, Wyrwich K, Wu A, Velikova G, Terwee C, Snyder C, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;**22**;1889–905. http://dx.doi.org/10.1007/s11136-012-0344-y

20. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, *et al.* Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *J Am Med Assoc* 2013;**309**:814–22. http://dx.doi.org/10.1001/jama.2013.879

21. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;**21**:651–7. http://dx.doi.org/10.1007/s11136-011-9960-1

22. Jenkinson C, Clarke C, Gray R, Hewitson P, Ives N, Morley D, *et al.* Comparing results from long and short form versions of the Parkinson's disease questionnaire in a longitudinal study. *Parkinsonism Relat Disord* 2015;**21**:1312–16. http://dx.doi.org/10.1016/j.parkreldis.2015.09.008

23. Coste J, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. *J Clin Epidemiol* 1997;**50**:247–52. http://dx.doi.org/10.1016/S0895-4356(96)00363-0

24. Juniper EF, Guyatt GH, Streiner DL, King DR. Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* 1997;**50**:233–8. http://dx.doi.org/10.1016/S0895-4356(96)00377-0

25. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010;**11**:215–25. http://dx.doi.org/10.1007/s10198-009-0168-z

26. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Methods for interpreting change over time in patient-reported outcome measures. Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). *Qual Life Res* 2013;**22**:475–83. http://dx.doi.org/10.1007/s11136-012-0175-x

27. Lloyd H, Jenkinson C, Hadi M, Gibbons E, Fitzpatrick R. Patient reports of the outcomes of treatment: a structured review of approaches. *Health Qual Life Out* 2014;**12**:5. http://dx.doi.org/10.1186/1477-7525-12-5

28. Black N, Varagunam M, Hutchings A. Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery. *BMJ Qual Saf* 2014;**23**:534–42. http://dx.doi.org/10.1136/bmjqs-2013-002707

29. Gibbons E, Hewitson P, Morley D, Jenkinson C, Fitzpatrick R. The Outcomes and Experiences Questionnaire: development and validation. *Patient Relat Outcome Meas* 2015;**6**:179–89. http://dx.doi.org/10.2147/PROM.S82784

30. Varagunam M, Hutchings A, Black N. Relationship between patient-reported outcomes of elective surgery and hospital and consultant volume. *Med Care* 2015;**53**:310–16. http://dx.doi.org/10.1097/mlr.0000000000000318

31. Varagunam M, Hutchings A, Black N. Do patient-reported outcomes offer a more sensitive method for comparing the outcomes of consultants than mortality? A multilevel analysis of routine data. *BMJ Qual Saf* 2015;**24**:195–202. http://dx.doi.org/10.1136/bmjqs-2014-003551

32. Hutchings A, Neuburger J, van der Meulen J, Black N. Estimating recruitment rates for routine use of patient reported outcome measures and the impact on provider comparisons. *BMC Health Serv Res* 2014;**14**:66. http://dx.doi.org/10.1186/1472-6963-14-66

33. Peters M, Crocker H, Jenkinson C, Doll H, Fitzpatrick R. The routine collection of patient-reported outcome measures (PROMs) for long-term conditions in primary care: a cohort survey. *BMJ Open* 2014;**4**:e003968. http://dx.doi.org/10.1136/bmjopen-2013-003968

34. Hazell ML, Morris JA, Linehan MF, Frank PI, Frank TL. Factors influencing the response to postal questionnaire surveys about respiratory symptoms. *Prim Care Respir J* 2009;**18**:165–70. http://dx.doi.org/10.3132/pcrj.2009.00001

35. Hutchings A, Neuburger J, Grosse Frie K, Black N, van der Meulen J. Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England. *Health Qual Life Outcomes* 2012;**10**:34. http://dx.doi.org/10.1186/1477-7525-10-34

36. Peters M, Crocker H, Dummett S, Jenkinson C, Doll H, Fitzpatrick R. Change in health status in long-term conditions over a one year period: a cohort survey using patient-reported outcome measures. *Health Qual Life Outcomes* 2014;**12**:123. http://dx.doi.org/10.1186/s12955-014-0123-2

37. Gomes M, Gutacker N, Bojke C, Street A. Addressing missing data in patient-reported outcome measures (PROMS): implications for the use of proms for comparing provider performance [published online ahead of print 5 March 2015]. *Health Econ* 2015. http://dx.doi.org/10.1002/hec.3173

38. Janssens A, Rogers M, Thompson Coon J, Allen K, Green C, Jenkinson C, *et al.* A systematic review of generic multidimensional patient-reported outcome measures for children, part I: descriptive characteristics. *Value Health* 2015;**18**:315–33. http://dx.doi.org/10.1016/j.jval.2014.12.006

39. Matza L, Patrick D, Riley A, Alexander J, Rajmi L, Pleil A, *et al.* Pediatric patient-reported outcome instruments for research to support medical product labeling: report of the ISPORPRO good research practices for the Assessment of Children and Adolescents Task Force. *Value Health* 2013;**16**:461–79. http://dx.doi.org/10.1016/j.jval.2013.04.004

40. Aguirre E, Kang S, Hoare Z, Edwards RT, Orrell M. How does the EQ-5D perform when measuring quality of life in dementia against two other dementia-specific outcome measures? *Qual Life Res* 2016;**25**:45–9. http://dx.doi.org/10.1007/s11136-015-1065-9

41. Grske J, Meyer S, Wolf-Ostermann K. Quality of life ratings in dementia care? A cross-sectional study to identify factors associated with proxy-ratings. *Health Qual Life Outcomes* 2014;**12**:177. http://dx.doi.org/10.1186/s12955-014-0177-1

42. Parsons N, Griffin X, Achten J, Costa M. Outcome assessment after hip fracture: is EQ-5D the answer? *Bone Joint Res* 2014;**3**:69–75. http://dx.doi.org/10.1302/2046-3758.33.2000250

43. Jahagirdar D, Kroll T, Ritchie K, Wyke S. Patient-reported outcome measures for chronic obstructive pulmonary disease: the exclusion of people with low literacy skills and learning disabilities. *Patient* 2013;**6**:11–21. http://dx.doi.org/10.1007/s40271-013-0004-5

44. Beadle-Brown J, Ryan S, Windle K, Holder J, Turnpenny A, Smith N, *et al. Engagement of People with Long Term Conditions in Health and Social Care Research: Barriers and Facilitators to Capturing the Views of Seldom-Heard Populations*. Quality and Outcomes of Person-centred Care Policy Research Unit; 2012. URL: www.pssru.ac.uk/archive/pdf/4390.pdf (accessed 10 April 2016).

45. Boyce M, Browne J, Greenhalgh J. Surgeon's experiences of receiving peer benchmarked feedback using patient-reported outcome measures: a qualitative study. *Implement Sci* 2014;**9**:84. http://dx.doi.org/10.1186/1748-5908-9-84

46. Bausewein C, Simon S, Benalia H, Downing J, Mwangi-Powell F, Daveson B, *et al.* Implementing patient reported outcome measures (PROMs) in palliative care – users' cry for help. *Health Qual Life Outcomes* 2011;**9**:27. http://dx.doi.org/10.1186/1477-7525-9-27

47. Gilbert A, Sebag-Montefiore D, Davidson S, Velikova G. Use of patient-reported outcomes to measure symptoms and health related quality of life in the clinic. *Gynecol Oncol* 2015;**136**:429–39. http://dx.doi.org/10.1016/j.ygyno.2014.11.071

48. Jensen RE, Rothrock NE, DeWitt EM, Spiegel B, Tucker CA, Crane HM, *et al.* The role of technical advances in the adoption and integration of patient-reported outcomes in clinical care. *Med Care* 2015;**53**:153–9. http://dx.doi.org/10.1097/MLR.0000000000000289

49. Santana M, Haverman L, Absolom K, Takeuchi E, Feeny D, Grootenhuis M, *et al.* Training clinicians in how to use patient-reported outcome measures in routine clinical practice. *Qual Life Res* 2015;**24**:1707–18. http://dx.doi.org/10.1007/s11136-014-0903-5

50. Wolpert M. Uses and abuses of patient reported outcome measures (PROMs): potential iatrogenic impact of PROMs implementation and how it can be mitigated. *Adm Policy Ment Health* 2014;**41**:141–5. http://dx.doi.org/10.1007/s10488-013-0509-1

51. Hildon Z, Neuburger J, Allwood D, van der Meulen J, Black N. Clinicians' and patients' views of metrics of change derived from patient reported outcome measures (PROMs) for comparing providers' performance of surgery. *BMC Health Serv Res* 2012;**12**:171. http://dx.doi.org/10.1186/1472-6963-12-171

52. Van der Wees P, Nijhuis Van der Sanden M, Ayanian G, Black N, Westert G, Schneider E. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. *Milbank Q* 2014;**92**:754–75. http://dx.doi.org/10.1111/1468-0009.12091

53. Wu AW, Kharrazi H, Boulware LE, Snyder CF. Measure once, cut twice – adding patient-reported outcome measures to the electronic health record for comparative effectiveness research. *J Clin Epidemiol* 2013;**66**(Suppl. 8):12–20. http://dx.doi.org/10.1016/j.jclinepi.2013.04.005

54. Warrington L, Absolom K, Velikova G. Integrated care pathways for cancer survivors – a role for patient-reported outcome measures and health informatics. *Acta Oncol* 2015;**54**:600–8. http://dx.doi.org/10.3109/0284186X.2014.995778

55. Nama V, Nordin A, Bryant A. Patient-reported outcome measures for follow-up after gynaecological cancer treatment. *Cochrane Database Syst Rev* 2013;**11**:CD010299. http://dx.doi.org/10.1002/14651858.cd010299.pub2

56. Kroenke K, Monahan P, Kean J. Pragmatic characteristics of patient-reported outcome measures are important for use in clinical practice. *J Clin Epidemiol* 2015;**68**:1085–92. http://dx.doi.org/10.1016/j.jclinepi.2015.03.023

Essay 5 Evaluating health-care equity

Rosalind Raine,[1] Zeynep Or,[2] Stephanie Prady[3] and Gywn Bevan[4]

[1]Department of Applied Health Research, University College London, London, UK
[2]Institut de Recherche et Documentation en Économie de la Santé, Paris, France
[3]Department of Health Sciences, University of York, York, UK
[4]Department of Management, London School of Economics and Political Science, London, UK

Declared competing interests of authors: none

Published May 2016
DOI: 10.3310/hsdr04160-69

This essay should be referenced as follows:

Raine R, Or Z, Prady S, Bevan G. Evaluating health-care equity. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 69–84.

List of abbreviations

CI	confidence interval
FNP	Family Nurse Partnership
GP	general practitioner
HR	hazard ratio
IMD	Index of Multiple Deprivation
PMB	post-menopausal bleeding
QCA	Qualitative Comparative Analysis
RAWP	Resource Allocation Working Party
SEC	socioeconomic circumstances

Abstract

The evaluation of health-care equity necessitates measuring both horizontal and vertical equity components to establish whether or not patients receive the health care across levels of need. Examining interactions between social factors of concern and need, or stratifying analyses according to different levels of need, can be used to identify horizontal and vertical equity. Increased data linkage across sectors and settings is vital for identification of sources of inequity and, crucially, to ascertain whether or not any identified variation has clinical relevance. Macro-, meso- and micro-level determinants of equity should always, ideally, be considered. However, examination of the 'gap' or the 'gradient' will depend on the intervention studied. Emerging techniques should be harnessed, for example machine learning to more completely exploit data sources on need, case mix and outcomes; and interactive multimedia techniques to examine social variations in clinical decision-making.

Scientific summary

Universal health-care systems aim to provide care for all, solely according to clinical need. Despite this, the 'inverse care law' has been demonstrated to operate in these systems. In this essay we discuss key methodological challenges in the measurement of health-care equity and propose ways forward.

The ascertainment of equity in health care requires the evaluation of both horizontal and vertical components: horizontal equity refers to the equal treatment of those with equal needs and vertical equity recognises that people with greater clinical needs should have more intervention. Methods to examine varying levels of need include stratifying analyses according to different levels of need or examining interactions between the social factor of concern and a measure of need.

Data on the population in need of care and case-mix data may be poorly recorded or available only as free text. Emerging innovations in machine learning enable extraction of valuable information from both existing and hitherto unused data sources. Data linkage of data from different settings is also required to identify source(s) of inequity and whether or not it matters in terms of outcomes.

Careful consideration should be given on whether it is most appropriate to examine the gap or the gradient in sociodemographic differences in the uptake of health-care/public health interventions. Statistical techniques are now available to calculate the sample size needed for gradient analyses.

In addition to examining equity at the intervention level, innovative methods are now being developed to examine the impact of national policies. Evidence from North America is that dismantling universal systems will not only create greater inequities in access but also mean increases in total costs of health care. Researching inequalities at the individual level and in clinical encounter are also beginning to be addressed. Recent approaches include the use of web-based interactive multimedia vignettes with actor 'patients' to simulate key features of health-care consultations. Qualitative, longitudinal study designs which allow the exploration of decision-making at different points of a patient's health-care journey could also yield valuable insights.

The development of innovative techniques and application of emerging methods for collecting and analysing data will enable equity to be measured with greater accuracy, precision, relevance and comprehensiveness. This will, in turn, better inform interventions for their remediation.

Introduction

Universal health-care systems, whether publicly funded or insurance based, aim to provide health care for all, according to clinical need, and undistorted by social or economic factors, geographical location or ability to pay. Yet observers have long recognised the presence of the so-called 'inverse care law' operating in these systems.[1] This term was first coined in 1971 by Dr Julian Tudor Hart, a general practitioner (GP) who worked in socially deprived mining communities in the Welsh valleys. His observation that 'the availability of good medical care tends to vary inversely with the need for it in the population served' was largely based on his personal experiences rather than on empirical research. However, a wealth of confirmatory data has since been published internationally. In England, this led the Chief Medical Officer to observe in 2005 that, in the publicly funded NHS:

> *. . . healthcare [. . .] is to some extent inequitable at present: the preference of clinicians, the socioeconomic status and empowerment of patients, and decisions regarding specific local resource allocation may influence clinical practice as much as the actual health needs of patients, the behaviour of any pathological process or the scientific evidence base.*
>
> *p. 22 (reproduced under the terms of the Open Government Licence for Public Sector Information)*[2]

For those health-care systems which aim to provide universal coverage based solely on clinical need, health-care inequity matters because it undermines the capacity of the system to remain true to its core values or its constitution. The robust measurement of equity is, therefore, crucial.

In this essay we discuss key methodological challenges in the measurement of health-care equity and propose some ways forward. We also briefly review evidence on policies to achieve equity using the method of the natural experiment.

Equality, equity and its horizontal and vertical components

Equity is about fairness and justice and implies that everyone should have an equal opportunity to attain their full potential for health or for the use of health care.[3] It should be distinguished from the related concept of equality. Equality is about the equal distribution of shares (of health or health care) so that each individual receives the same amount. The notion of equity transcends equality. Some inequalities may be unavoidable and therefore are not generally considered unjust. Others, for example associated with one's area of residency, ethnic group, sex, age, socioeconomic circumstances (SEC) or disability, might be avoided and so are considered inequitable. They may be unfair or unjust as well as unequal and, because they are not solely determined by need, they may lead to differences in outcomes.

We recognise that definitions of fairness vary according to libertarian, liberal and collectivist perspectives. However, universal health-care systems define fairness in terms of needs. Mooney[4] pointed out that equitable distribution according to clinical need might refer to the distribution of expenditure, access, use or health. With respect to expenditure distribution, in the UK, the NHS uses formulae to promote equitable allocation of funding for care. The Resource Allocation Working Party (RAWP) and its successors have developed different methods of formula funding for the four countries of the UK (England, Scotland, Northern Ireland and Wales) which recommend that money should be distributed on the basis of population size, weighted for relative need and accounting for variations in unavoidable costs of providers (e.g. higher living costs in London, excess costs for delivering services in remote areas in Scotland).[5] The components of the RAWP formula have been reviewed and revised for over 30 years and they continue to be subject to reanalysis. For example, although the RAWP recommendation to account for relative need by weighting for age and sex using national average rates of utilisation is relatively uncontroversial, their use of the standardised mortality ratio to account for additional need has raised many questions, which are discussed by Bevan.[5] In 1999, a new objective for the allocation of resources in the English NHS was introduced to contribute to the reduction in avoidable health inequalities.[6] As a consequence, a health inequalities component was introduced into the allocation formula and increases in allocations have since favoured more deprived areas. Local NHS commissioners were free to use these additional funds to purchase primary or secondary health care or public health services, to better meet their population's needs and improve the quality of care received. The underlying rationale is that additional health-care expenditure translates into improved population health outcomes. A recent analysis of this policy found that geographical inequalities in mortality from causes amenable to health care declined in absolute terms during the 10-year period in which this allocation policy was applied.[7] Moreover, the association between additional NHS funds and reduced mortality was stronger in deprived areas than in more affluent areas. Analyses such as these, using aggregated, routinely available data, cannot exclude the possibility that the associations observed were attributable to unmeasured confounders, such as smoking, or to other, concurrent non-NHS policies implemented to tackle social exclusion in disadvantaged areas. Moreover, they do not tell us about the types and content of services received and their relative contributions to outcomes.

Indicators of access include the availability of resources, waiting times, user charges and others barriers to care. Thus, equity of access is a purely supply-side phenomenon, in the sense that equal services are made available to patients in equal need. The difficulty with measuring equity according to this definition is that individuals may not use services to which they have access. Their reasons for non-use are socially patterned and are, at least in part, because of structural and environmental barriers which are proportionately greater for the socioeconomically disadvantaged, older people, people with disabilities or those for whom English is not their first language. Routine data such as medical records may not, therefore, strictly give a measure of access, because they can supply information only about services which have been taken up.

The definition of equity which takes this into account is equal use for equal need. This definition recognises the influence of both demand and supply factors on the pursuit of equity. These factors are influenced by the preferences, perceptions and prejudices of both patient and health-care provider. Most studies on equity and access analyse utilisation rates (often adjusted by sociodemographic and clinical indicators of need) as a proxy measure of access.

The final definition is equal health for equal need. This definition addresses the fact that there are inequalities in health arising from the level of resources, housing conditions, exposure to environmental hazards and different lifestyles and behaviours. It is an ideal rather than an operational definition, as both avoidable and unavoidable influences affect our health. Neighbourhood renewal[8] and Sure Start[9] are examples of initiatives in the UK that move towards this ideal.

There is another issue surrounding distribution according to need which should be considered. This is that distribution on the basis of need comes in two versions: a horizontal version (people with equal needs should be treated the same) and a vertical version [people with greater clinical needs should have more intervention (provided it is effective) than those with lesser needs (unequal use for unequal need)]. From a macroeconomic perspective, Mooney and Jan[10] defined vertical equity in terms of positive discrimination, arguing that achievement of horizontal equity is rarely enough. They propose differential weighting of the level of need by socioeconomic group in order to more thoroughly investigate the distribution of health and health care and to facilitate the movement of disadvantaged people up to the level of the advantaged.[10] Mooney and Jan[10] do not provide the differential weights that would allow examination of the presence of vertical equity. Sutton[11] takes the approach forward by specifying a target level of use for people at different levels of need and then comparing whether or not actual use equals the target.

An alternative way of operationalising vertical equity is generated by the understanding that the demonstration of equal use for equal need does not necessarily indicate unequal use for unequal need.[12] For example, although male and female patients with a mild form of a disease may be treated equally (horizontal equity), it cannot be assumed that the likelihood of treatment varies according to the degree of abnormality in both men and women. Men with severe disease may be more likely to receive treatment than men with mild disease (vertical equity), but the likelihood of treatment for women may not differ with disease severity (vertical inequity). The vertical component is often overlooked by researchers and policy-makers, and this prevents the comprehensive measurement of equity, the likely consequence of which is overestimation of fair use of care. Unless both components of equity are measured, it cannot be concluded that patients are receiving the health care that they need. Studies that use multivariable analysis alone to adjust for need assume that social differences in use are the same at every level of need, which may not be the case. Other analysis strategies should be employed to examine varying levels of need, such as stratifying analyses according to different levels of need or examining interactions between the social factor of concern and a measure of need.[12]

Identifying the population in need of care

To examine the equitable distribution of health care and public health interventions, we need to understand how interventions are distributed across the social factors of concern. This requires identification of both the complete population in need and their sociodemographic characteristics.

This is not always straightforward. Many of the social factors of interest, such as ethnicity, language skill or religion, are poorly recorded in national data sets or available only as free text (which is unstandardised). These data may not always be missing at random, potentially leading to selection bias. An emerging innovation is to apply machine learning techniques to 'code' free-text data in medical records. Thoughtful application of these methods has the potential to increase our ability to extract valuable information from both existing and hitherto unused or underused data sources. Machine learning could revolutionise systematic data collection in terms of comprehensiveness, completeness, timeliness and accuracy.

There are many sources of data that can be used to characterise socioeconomically disadvantaged groups. Area-based measures of SEC are often freely available and readily linked with postcode data; for example, in England the Index of Multiple Deprivation (IMD) scores and ranks areas containing approximately 650 households across several domains of deprivation.[13] Relying on area-based measures to ascertain SEC for individuals, however, can result in ecological fallacy, where lack of precision or incorrect inference stems from within-area pockets of relative advantage and disadvantage masked by the 'average' area score. In the absence of individual-level data, such effects can potentially be reduced by adding in data from commercially available data sets at smaller geographic units which provide information on consumer habits. Such data may distinguish the advantaged from the disadvantaged within an area and provide greater resolution on SEC when used in combination with the IMD.[14] The method is not bias free, however, as relatively high levels of consumerism may be the result of high debt, not affluence. Health and health-care researchers are perhaps less familiar with using sources such as local and national taxation information and education data, but forging links with data providers to explore such sources has the potential to reap benefits for future equity research.

Some relationships between health and measures of SEC are best characterised by J-shaped curves, meaning that the most disadvantaged group may not be where you expect to find them. For example, in the USA, health insurance coverage may not be linear with income; the wealthy may opt not to have health insurance, and benefit recipients may have state-provided coverage. Policy-related changes leading to changes in insurance patterns for different social groups over time complicate longitudinal analyses. The relationship between coverage and health for working poor is likely to be highly variable and associated at the micro level with heterogeneous contextual factors such as size of employer and family composition. In England, benefit recipients (and those with long-term chronic conditions) are exempt from prescription charges, so in this context it may be the working poor paying for prescriptions who suffer the highest financial burden of ill health. Examining and characterising the group most at risk by considering the context of the population setting is a crucial step in measuring inequalities. Quadratic terms can be added to regression models to allow for non-linear relationships.[15]

Accurate ascertainment of the population in need of health care may be hampered by several factors. First, incidence and prevalence from health-care data are likely to be inaccurate if there are a high number of undiagnosed cases of the condition in the community. Diagnosis (and diagnostic delays) are influenced by help-seeking behaviour, which is itself socially patterned. Second, disease presentation can also vary by social group. For example, women with ischaemic heart disease may present with symptoms that differ from the typical chest pain presentation by men.[16] This could lead to underascertainment in women if these differences are not reflected in study diagnosis criteria. Third, there may be a lack of consensus about the definition of need for intervention, for example in total hip replacement. Where the clinical rationale for intervention is not clear-cut, investigation and treatment may be influenced as much by the availability of doctors and diagnostic equipment, or by financial factors (such as a fee-for-service system), as by clinical need.[17] This highlights the problem of inequity owing to the overuse of medical interventions, which puts patients at risk of complications unnecessarily and drives up the cost of health care. Finally, health-care systems which are not universal are limited by their access to medical record databases which relate only to the insured population, excluding those without coverage, and there may be also unknown missing data bias for out-of-area medical visits.

Examining whether or not inequalities matter

We have already described why it is necessary to examine whether or not health-care inequalities matter in terms of their impact on outcomes. For example, in a study of the effect of secondary prevention in 30-day stroke survivors it was found that people aged 80–89 years were only half as likely to receive a lipid-lowering drug as those aged 50–59 years.[18] This treatment inequality was important because the receipt of secondary drug prevention was associated with a halving of the mortality risk. Crucially, there was little evidence that the effect of treatment differed by age. Therefore, the undertreatment of older people cannot be justified, unless it is explained by informed patient choice. If patient preferences explain some of the differences observed, then it is important to unravel their origin. This is a theme that we discuss later, in *Social variations in the clinical encounter*.

The gap versus the gradient

Policy-makers distinguish between the *gap* (the relative difference between advantaged and disadvantaged groups) in service provision, and the *gradient* (the continuum along which increasingly worse health is associated with a unit drop in the social factor of interest). An example of an approach to target the most disadvantaged subgroups only is the Nurse Family Partnership in the USA,[19] named the Family Nurse Partnership (FNP) in the UK[20] and VoorZorg in the Netherlands,[21] which provides intensive support to multiply disadvantaged first-time mothers. Three evaluations of these programmes in the USA report positive results for mothers and their babies, including improvements in birth outcomes, children's cognitive development and uptake of preventative health care.[19,22,23] A trial of the intervention in the

Netherlands reported lower smoking rates in nurse-visited pregnant women and increased breastfeeding duration. However, in England, where FNP was added to the usually provided health and social care, no additional benefits were achieved in the primary outcomes, which included smoking in pregnancy, birthweight, rates of second pregnancies and emergency hospital visits for the child.[24] There were important differences in both the trial design and the health-care context between the US and English studies: the English trial was a large, pragmatic, independently led evaluation, in contrast to the US evaluations, which were single centre and led by the intervention developers with a greater emphasis on efficacy. Women in England have access to more statutory public health, health care and social services than US women. The lack of any additional benefit from the English FNP suggests that macroeconomic policies and structural changes are required to complement intensive support services when tackling complex, multifactorial and enduring problems.

The costs of inequalities are borne not only by those at the bottom of the socioeconomic hierarchy, but by those at every level. Policies that target the most disadvantaged subgroups only, or which aim to narrow the gap between the most and least disadvantaged, underestimate the pervasive effect across the socioeconomic hierarchy and exclude those in need in the intermediate socioeconomic groups. Even for targeted interventions which are found to be effective, the population-level impact maybe smaller for these than for universal interventions. Such arguments augur in favour of tackling the gradient to address inequalities. Although it is rarely done, the gradient can be measured to examine the effect of universal interventions which are expected to reach the whole population, for example in Wardle *et al.*[25] Where gradients are measured, the time frame over which effects are evaluated should be carefully chosen. This is because universal interventions can increase health inequalities in the short term, as the more advantaged are usually the first to take up new services, but this effect can level over time as the advantaged groups plateau in terms of their ability to benefit.[26]

Calculating the sample size needed for gradient analyses can present challenges. One strategy was developed for a universal screening programme intervention in which uptake needed to increase more in disadvantaged groups than in advantaged groups.[25] Here, the solution was to use the weighted averages of the association between the deprivation quintile and the previously observed response to screening instead of the usual proportions in the formula, with the response rate held a constant across quintiles.[27] Efficiency can be optimised by post-stratification, where treatment groups are stratified with a pre-treatment variable, treatment effects within the strata are estimated, and the weighted average of these estimates is used to calculate the overall average treatment effect estimate.[28] In studies that aim to estimate average effects, reporting by disadvantaged subgroup, even with insufficient within-study power, increases the pool of health inequality studies available for potential synthesis. There are also additional approaches that can be taken alongside subgroup analyses to estimate within-study differential effects. One method is to conduct a latent class analysis across the whole sample where key response patterns are identified across different social groups of participants.[29] This method has the advantage of not being sample size dependent, although very small samples with many multicategory variables may run into estimation problems.

Identifying the sources of inequitable provision of care

Most research undertaken in this area examines sociodemographic variations in health-care use for a defined intervention (or package of interventions) at one point in the management pathway. Inequalities have been demonstrated at each stage of the pathway: in participation in population-based screening programmes in the community, in the management of health problems in primary care, in the access to and use of diagnostic and therapeutic procedures within secondary care, and in rehabilitation and end-of-life care.

For example, the UK breast, cervical and bowel cancer screening programmes are run by the NHS without financial cost to participants. Nonetheless, the uptake of all programmes shows a gradient by SEC.[30,31]

The strongest gradient is for bowel cancer screening. This involves offering a guaiac faecal occult blood testing kit for use at home. In the first 2.6 million invitations in 2006–9, uptake was 61% in the least deprived quintile of residential areas and only 35% in the most deprived quintile.[32] Bowel cancer screening is currently being extended to include one-off flexible sigmoidoscopy of the lower bowel to identify polyps that may develop into bowel cancer. Overall uptake in the first six pilot centres was 33% in the most deprived areas, rising to 53% in the most affluent areas.[33]

Within primary care, Lyratzopoulos *et al.*[34] demonstrated that, in England, younger patients, ethnic minorities and women are more likely to have visited their GP a minimum of three times prior to referral to hospital for cancer diagnosis, suggesting potential avoidable delay in their management.

Sociodemographic variations in the likelihood of referral by GPs into secondary care has also been found to vary depending on the presence of explicit national guidelines on referral.[35] This study used The Health Improvement Network, a widely used primary care database, to examine sociodemographic variations in referral for potentially life-threatening conditions where national guidance on referral has been published [post-menopausal bleeding (PMB) and dyspepsia in people > 55 years of age] and for symptoms where there is clinical uncertainty regarding the decision to refer (hip pain and dyspepsia in people < 55 years of age). For the three conditions examined, older patients were less likely to be referred, after adjusting for comorbidity. Women were less likely than men to be referred for hip pain [hazard ratio (HR) 0.90, 95% confidence interval (CI) 0.84 to 0.96]. More deprived patients with hip pain and dyspepsia (if < 55 years old) were less likely to be referred. Adjusted HRs for those in the most deprived quintile compared with the least deprived were 0.72 (95% CI 0.62 to 0.82) and 0.76 (95% CI 0.68 to 0.85), respectively. There was no socioeconomic gradient in referral for PMB. These findings are important given the widespread prevalence of non-specific symptoms for which explicit referral guidance does not exist, but which could, nonetheless, be indicative of serious underlying pathology (e.g. lung, colorectal and ovarian cancers).

Access to timely secondary care has also been demonstrated to be inequitable: in England, patients from deprived areas, older people and women are more likely to be admitted to hospital as emergencies than electives for colorectal, breast and lung cancer. This was inequitable (rather than simply unequal) because people from deprived areas and older people were also less likely to receive preferred surgical procedures such as breast-conserving surgery and lung cancer resection.[36] The research was limited by the data available from routinely collected Hospital Episode Statistics. The findings suggest that social variations in both timely presentation and pathways to care need to be prospectively examined. Prospective data collection would allow the impact of potential confounders, such as tumour characteristics including stage, as well as case mix and patient preferences, to be examined.

Once in the secondary care system, women from less affluent neighbourhoods in France have lower probability of receiving 'best practice' treatments (Or Z, Rococo E, Bonastre J, Institute for Research and Information in Health Economics, France, 2016), and in England, patients referred from deprived practices have reduced levels of diagnostic angiography and higher waiting times.[37] There is also significant variation for women with breast cancer who live in deprived areas in England: they are more likely to be diagnosed at end stage (Stage IV)[38] and less likely to receive surgery and radiotherapy,[39] and area-based disparity in receipt of treatment may be the major driver of variation in lung cancer survival in England.[40] Inequalities in the effective provision of rehabilitation care are also evident.[41]

When inequalities such as these are uncovered it is not possible to conclude whether the disparity occurs at the level of the intervention under consideration, or as a consequence of inequalities in the provision of preceding interventions, or in direct (and appropriate) response to the results of previous investigations. For example, a review of sex bias in use of cardiac care found that high-quality prospective studies reported sex differences in favour of men in the use of angiography.[12] However, there was consistent evidence of no sex difference in those patients in whom the results of previous investigations had been taken into account. This indicates that the sex inequalities identified were not *inequitable*, in that they were fair and made on the basis of clinical need (identified by the earlier investigation). Thus, the entire management pathway needs to be examined to establish the reasons for the differences found. The challenge is to identify data that will permit unbiased ascertainment of need and use across two or more settings (primary, secondary and community care). It is here that carefully specified individual record linkage between survey data to databanks of routinely collected data, and linking different sources of routinely collected data, is of high value.

Furthermore, while attention to the meso (organisational) level is appropriate, a comprehensive understanding of the sources of inequity also requires consideration of the impact of national policies (i.e. macro-level factors).[42] For example, although the European Union provides universal coverage, the comprehensiveness of care varies from country to country. This is in part explained by the density and position of generalist clinicians, the presence of out-of-pocket payments and the referral system (gatekeeping or not), all of which have an impact on the use of preventative, primary and specialist care.[43] In addition, in social insurance-based systems, for example in France, complex rules for reimbursement for different services appear to act as a disincentive for service use among some population groups.[44]

In the UK, Cookson *et al.*[45] evaluated these policies by examining trends in primary care access, quality and outcomes by SEC between 2004 and 2012. They did this using routinely available whole population data including health data from four national administrative databases. The IMD (IMD 2010)[46] was used to assign socioeconomic status to neighbourhoods of approximately 1500 people each. Slope indices of inequality were measured in four indicators: patients per family doctor, primary care quality, preventable hospitalisation and amenable mortality. They found that, during this period, the NHS succeeded in substantially reducing socioeconomic inequalities in primary care access and quality but made only modest reductions in health-care outcome inequalities.

Such analyses cannot assess the extent to which observed trends in preventable hospitalisation and amenable mortality are attributable to, for example, trends in multimorbidity outside the control of the NHS or in social variations in illness behaviour. However, they both provide much-needed evidence of the influence of national policies and highlight the need to address health-care inequity in every sector and setting.

Finally, synthesising equity effects of health-care interventions can be challenging. One approach that may be useful when comparing several different interventions is Qualitative Comparative Analysis (QCA),[47] which can be used in tandem with quantitative analyses to work out what works for whom and under what conditions,[48] very much under a realist perspective.[49] Blackman and Dunstan[50] illustrate QCA's utility in health inequalities research by applying the method to survey data in order to understand the influence of place-based contextual factors related to variation in narrowing mortality gaps. Individual participant data meta-analyses (e.g. Virtanen *et al.*[51]) or other synthesis techniques such as QCA (e.g. Thomas *et al.*[52]) could be applied to studies in the health-care setting to explore variation in outcomes with regard to contextual effects such as the policy environment, and differential effects by disadvantage. The role of evidence synthesis in health services research is further discussed in *Essay 1* of this volume.

Social variations in the clinical encounter

A potential contributing cause of demonstrated health inequalities are social variations in individual behaviour and interactions between patients and health-care professionals. However, there are significant methodological challenges to researching the clinical encounter. Direct observation of doctors and patients offers no opportunity to control patients' clinical and sociodemographic characteristics, and would require observation of prohibitively vast numbers of consultations to obtain the necessary numbers in specific risk or demographic categories. Use of 'standardised' patients (i.e. consultations with doctors by trained actors) is considered a gold-standard method because it enables more control over patient characteristics, but it is costly. The use of fictional patient profiles (vignettes) can provide a valid, generalisable and efficient approach to studying variations in decision-making by health-care professionals. Most studies, however, use text-based vignettes, and omit many features of real consultations such as real-time responses to clinicians' questions or nuanced presentation of patient characteristics. This risks bias by offering clinicians a limited selection of response options, which primes them to consider certain actions. Many studies are small, with limited generalisability.

In one novel study, a website was constructed using interactive multimedia vignettes with actor 'patients' to simulate key features of consultations. GPs undertook consultations from each of six clinical profiles which varied according to the 'patients' sociodemographic characteristics and lung cancer risk. No GP saw the same actor twice. Within this constraint, allocation of GPs to vignettes was random. This achieved balance by sex, ethnicity and SEC, and thus GPs' decisions to initiate lung cancer investigation could be studied in a factorial design across different combinations of clinical and sociodemographic characteristics (Sheringham J, Sequeira R, Myles J, Hamilton W, McDonnell J, Offman J, *et al.*, University College London, 2016). This research demonstrated that, regardless of clinical risk, GPs were less likely to investigate older and black 'patients'.

Patient vignettes can provide insight into clinical decision-making but do not contribute towards our understanding of the patient perspective or of factors influencing the interaction between patients and health-care professionals. It is often suggested that patient choice underlies inequalities in uptake of health care and public health interventions. Even if decisions to forego some aspects of care reflect 'informed' choice, we still need to understand the origins of and influences on these choices. Choices are influenced by individual-level factors (motivation, perceived consequences of different actions and values placed on those consequences), social networks (families, peer groups), local environments (e.g. school ethos) and the national context (taxation, regulation, advertising, subsidies). There appear to be systematic differences in expectations for good health and in perceptions of risk and benefits of treatment between advantaged and less advantaged groups.[53] Patients' 'choice' for less intensive treatment may reflect inaccurate perceptions about the availability, effectiveness or risk of treatment, or be influenced by accurate observations that outcomes are worse in their community.[53] Furthermore, communication in clinical encounters is socially patterned, with clinicians providing less information and adopting a less participatory consulting style when consulting with less articulate, socially disadvantaged patients.[54] This is likely to directly impact on the decisions that patients make.[55]

Innovative approaches are therefore needed to explore and explain patient 'choice'. In-depth, observational, qualitative, longitudinal study designs are likely to be required which allow the exploration of decision-making at different points of a patient's health-care journey. For example, interviews before and after consultations about beliefs and expectations, and how they change, together with non-participant observation of clinical encounters, could yield valuable insights.

Designing interventions to improve equity

Given (and in spite of) the current pre-eminence of the concept of individual choice and responsibility, these observations suggest three levels of intervention. At the level of the patient–clinician interaction it is important to understand patients' and health professionals' assumptions about risks and benefits of medical interventions and their accuracy. Health-care delivery preferences may be shaped by quality of service, so we need to ensure equity in the quality of services and focus on quality improvement where certain communities are poorly served. At the societal level, the observation that preferences/perceptions of opportunities are often defined by social, economic and cultural factors underscores the importance of addressing fundamental structural inequities and sociocultural norms.

One area emerging in response to the problem of complex policy and population settings is to design intervention strategies on multiple levels. A host of interventions can often be theorised; for example, simultaneous economic, organisational and behavioural interventions may be needed to tackle diabetes.

The intervention framework needs consider factors at macro, meso and micro levels and their interactions, even though much of the dynamism may be focused at only one of these levels. Hawe[56] suggests ways to move towards interventions characterised not by a programme but by relationships, routines, power structures and sets of values. A number of essays in this volume take forward in a number of different ways the issue of complexity and how it can be appropriately addressed in evaluative research (see particularly *Essays 1*, *6* and *7*).

Can austerity threaten equity?

Cutler[57] argued that consequences of governments aiming for equity by increasing coverage were increases in costs, which led to top-down policies to control costs, which, in turn, resulted in a deterioration in quality (such as long waiting times), and the interest in market-type reforms to remedy that problem. Tuohy[58] has argued that governments' search for a system that achieves equity, controls costs with high-quality care and results in policy cycling as governments emphasise policies that tackle the most serious failure of these three objectives and implicitly neglect the other two. So, following the global financial crisis, fiscal pressures mean that governments now focus on cost control rather than on equity and quality. An obvious question is whether or not there is evidence that systems of universal coverage that are mainly free at the point of delivery, on grounds of promoting equity of access, are poorly designed to control costs. This essay has demonstrated that even with these structural characteristics, there is evidence that the inverse care law[1] still applies. Changing to partial coverage and high user charges would mean that that law would apply with even greater force. But the question is whether or not austerity would justify cycling to such policies as a way of controlling costs.

In trying to assess the impacts of government policies for the twin objectives of equity and cost control, we typically lack controlled experiments. Although there is the famous landmark study of the RAND Health Insurance Experiment,[59] which randomly allocated people to different insurance packages (with differing co-payments and co-insurance) and also included a Health Maintenance Organisation, this still falls short of assessing the macro questions of whether or not a financial system with universal coverage and services free at the point of delivery will have more serious problems of cost control than one with partial coverage and high user charges. Evans *et al.*[60–62] used the 'natural experiment' between the USA and Canada to investigate the impact of policies to improve equity of access to health care in terms of control of total costs. Prior to 1970, both countries were structurally similar in how health care was financed (with multiple insurers, partial coverage and high user charges) and in the delivery of care (with hospitals and physicians being independent of government and paid charges and fees for services delivered). After 1970, in Canada, only the financial system changed with the introduction of universal coverage free at the point of access for hospital care and physicians' services. Evans[60] emphasises that an unintended outcome of policy cycling by the Canadian government was that these policies that were directed at equity were later found to be highly effective in controlling total costs of health care in Canada as compared with the USA.

Evidence from 'natural experiments' is always open to challenge from potential influence from other factors and other outcomes that have not been measured. Their value is hence even more strongly dependent on developing a sound theoretical explanation, which is just what Evans[60] does in explaining what is a paradoxical outcome. Given the characteristics of health care, it is folly to regard patients as well-informed consumers making cost-conscious choices when confronted with high user charges, which are a poor way of trying to control costs when doctors typically make decisions. Indeed, Brook *et al.*[63] summarised the results of the RAND study as showing that '. . . cost sharing can be a blunt tool. It reduced both needed and unneeded health services' (p. 4)[63] and that '. . . subsequent RAND work on appropriateness of care found that economic incentives by themselves do not improve appropriateness of care or lead to clinically sensible reductions in service use' (p. 4).[63] Evans[60] goes on to argue that the key to cost control is targeting not patients but suppliers, and that is most effectively done by empowering government as the single payer with monopsony power in negotiations with physicians and hospitals.

Thus, he concludes: 'The standard theoretical analyses of health insurance, focussing on (an incomplete specification of) the incentives faced by patients, counts the peanuts but ignores the elephants'.[60]

Spend on health care as a percentage of Gross Domestic Product in 2013 was 16.4% for the USA, 10.2% for Canada and 8.4% for the UK.[64] Thus, although Canada and the USA have similar delivery systems, what appears to matter in controlling costs is that Canada and the UK have similar institutional financial systems. So, although austerity strains universal systems which are largely free at the point of delivery, the evidence from North America is that dismantling that system will not only create greater inequities in access but also mean increases in total costs of health care. Such universal systems need to be developed to manage insurance to relate to differences in patients' preferences (as revealed, e.g., through shared decision-making) but that is beyond the scope of this essay.

Conclusions

In this essay we have highlighted strategies to address some of the methodological challenges faced when evaluating health-care equity. These include the measurement of both horizontal and vertical equity, the use of machine learning to enhance the completeness and quality of data collection and considerations on the appropriateness of gap or gradient analyses. Linkages between routinely collected data from primary, hospital and social care; disease cohort and audit data, deprivation indices, mortality registers; and judiciously chosen other sources (e.g. from education and consumer surveys) increase our ability to precisely identify sources of inequity along the patient pathway and to disentangle the influences of case mix, social and organisational context. Together with innovative methods to capture expectations and patient and health professional decision-making in real time, these strategies will inform the design and evaluation of national-, organisational- and individual-level interventions to improve health-care equity.

Acknowledgements

The authors thank Professor David Byrne for his helpful comments.

Contributions of authors

Rosalind Raine (Professor, Applied Health Research) wrote the first draft of the essay.

Zeynep Or (Research Director, Health Economist) and **Gwyn Bevan** (Policy Analyst) provided additional material and along with **Stephanie Prady** (Research Fellow, Social Epidemiologist) commented on the draft.

References

1. Hart JT. The inverse care law. *Lancet* 1971;**1**:405–12. http://dx.doi.org/10.1016/S0140-6736(71)92410-X
2. Department of Health. *The Chief Medical Officer on the State of Public Health: Annual Report 2005*. London: Department of Health; 2006. URL: http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4137367.pdf (accessed 10 April 2016).
3. Roberts T. What is the difference between equity and equality? *J Health Serv Res* 1997;**2**:129.
4. Mooney GH. Equity in health care: confronting the confusion. *Eff Health Care* 1983;**1**:179–85.
5. Bevan G. The search for a proportionate care law by formula funding in the English NHS. *Finance Account Manag* 2009;**25**:391–410. http://dx.doi.org/10.1111/j.1468-0408.2009.00484.x

6. Department of Health. *The NHS Plan: A Plan for Investment, A Plan for Reform*. London: The Stationery Office; 2000.

7. Barr B, Bambra C, Whitehead M. The impact of NHS resource allocation policy on health inequalities in England 2001–11: longitudinal ecological study. *BMJ* 2014;**348**:g3231. http://dx.doi.org/10.1136/bmj.g3231

8. Social Exclusion Unit, Cabinet Office. *A New Commitment to Neighbourhood Renewal – National Strategy Action Plan*. London: Social Exclusion Unit, Cabinet Office; 2001.

9. Belsky J, Barnes J, Melhuish EC. *The National Evaluation of Sure Start: Does Area-Based Early Intervention Work?* Bristol: The Policy Press; 2007.

10. Mooney G, Jan S. Vertical equity: weighting outcomes? Or establishing procedures? *Health Policy* 1997;**39**:79–87. http://dx.doi.org/10.1016/S0168-8510(96)00851-2

11. Sutton M. Vertical and horizontal aspects of socio-economic inequity in general practitioner contacts in Scotland. *Health Econ* 2002;**11**:537–49. http://dx.doi.org/10.1002/hec.752

12. Raine R. Bias measuring bias. *J Health Serv Res Policy* 2002;**7**:65–7. http://dx.doi.org/10.1258/1355819021927584

13. Noble M, McLennan D, Wilkinson K, Withworth A, Barnes H. *The English Indices of Deprivation* 2007. London: Communities and Local Government; 2008.

14. Sheringham J, Sowden S, Stafford M, Simms I, Raine R. Monitoring inequalities in the National Chlamydia Screening Programme in England: added value of ACORN, a commercial geodemographic classification tool. *Sex Health* 2009;**6**:57–62. http://dx.doi.org/10.1071/SH08036

15. Gray L, Leyland AH. A multilevel analysis of diet and socio-economic status in Scotland: investigating the ‘Glasgow effect’. *Public Health Nutr* 2009;**12**:1351–8. http://dx.doi.org/10.1017/S1368980008004047

16. Brewer LC, Svatikova A, Mulvagh SL. The challenges of prevention, diagnosis and treatment of ischemic heart disease in women. *Cardiovasc Drugs Ther* 2015;**29**:355–68. http://dx.doi.org/10.1007/s10557-015-6607-4

17. Welch WP, Miller ME, Welch HG, Fisher ES, Wennberg JE. Geographic variation in expenditures for physicians’ services in the United States. *N Engl J Med* 1993;**328**:621–7. http://dx.doi.org/10.1056/NEJM199303043280906

18. Raine R, Wong W, Ambler G, Hardoon S, Petersen I, Morris R, *et al.* Sociodemographic variations in the contribution of secondary drug prevention to stroke survival at middle and older ages: cohort study. *BMJ* 2009;**338**:b1279. http://dx.doi.org/10.1136/bmj.b1279

19. Olds DL. Prenatal and infancy home visiting by nurses: from randomized trials to community replication. *Prev Sci* 2002;**3**:153–72. http://dx.doi.org/10.1023/A:1019990432161

20. FNP. *Family Nurse Partnership*. 2015. URL: http://fnp.nhs.uk/ (accessed 10 April 2016).

21. Mejdoubi J, van den Heijkant SC, van Leerdam FJ, Crone M, Crijnen A, HiraSing RA. Effects of nurse home visitation on cigarette smoking, pregnancy outcomes and breastfeeding: a randomized controlled trial. *Midwifery* 2014;**30**:688–95. http://dx.doi.org/10.1016/j.midw.2013.08.006

22. Olds DL, Henderson CR Jr, Chamberlin R, Tatelbaum R. Preventing child abuse and neglect: a randomized trial of nurse home visitation. *Pediatrics* 1986;**78**:65–78.

23. Kitzman H, Olds DL, Henderson CR Jr, Hanks C, Cole R, Tatelbaum R, *et al.* Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. A randomized controlled trial. *JAMA* 1997;**278**:644–52. http://dx.doi.org/10.1001/jama.1997.03550080054039

24. Robling M, Bekkers MJ, Bell K, Butler CC, Cannings-John R, Channon S, *et al.* Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (building blocks): a pragmatic randomised controlled trial. *Lancet* 2016;**387**:146–55. http://dx.doi.org/10.1016/S0140-6736(15)00392-X

25. Wardle J, von Wagner C, Kralj-Hans I, Halloran SP, Smith SG, McGregor LM, *et al.* Effects of evidence-based strategies to reduce the socioeconomic gradient of uptake in the English NHS Bowel Cancer Screening Programme (ASCEND): four cluster-randomised controlled trials. *Lancet* 2016;**387**:751–9. http://dx.doi.org/10.1016/S0140-6736(15)01154-X

26. Victora CG, Vaughan JP, Barros FC, Silva AC, Tomasi E. Explaining trends in inequities: evidence from Brazilian child health studies. *Lancet* 2000;**356**:1093–8. http://dx.doi.org/10.1016/S0140-6736(00)02741-0

27. Brentnall AR, Duffy SW, Baio G, Raine R. Strategy for power calculation for interactions: application to a trial of interventions to improve uptake of bowel cancer screening. *Contemp Clin Trials* 2012;**33**:213–17. http://dx.doi.org/10.1016/j.cct.2011.09.021

28. Miratrix LW, Sekhon JS, Yu B. Adjusting treatment effect estimates by post-stratification in randomized experiments. *J Roy Stat Soc B Met* 2013;**75**:369–96. http://dx.doi.org/10.1111/j.1467-9868.2012.01048.x

29. Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevent Sci* 2013;**14**:157–68. http://dx.doi.org/10.1007/s11121-011-0201-1

30. Maheswaran R, Pearson T, Jordan H, Black D. Socioeconomic deprivation, travel distance, location of service, and uptake of breast cancer screening in North Derbyshire, UK. *J Epidemiol Community Health* 2006;**60**:208–12. http://dx.doi.org/10.1136/jech.200X.038398

31. Bang JY, Yadegarfar G, Soljak M, Majeed A. Primary care factors associated with cervical screening coverage in England. *J Public Health (Oxf)* 2012;**34**:532–8. http://dx.doi.org/10.1093/pubmed/fds019

32. Logan RF, Patnick J, Nickerson C, Coleman L, Rutter MD, von Wagner C. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut* 2012;**61**:1439–46. http://dx.doi.org/10.1136/gutjnl-2011-300843

33. McGregor LM, Bonello B, Kerrison RS, Nickerson C, Baio G, Berkman L, *et al.* Uptake of bowel scope (flexible sigmoidoscopy) screening in the English national programme: the first 14 months [published online ahead of print 20 September 2015]. *J Med Screen* 2015. http://dx.doi.org/10.1177/0969141315604659

34. Lyratzopoulos G, Neal RD, Barbiere JM, Rubin GP, Abel GA. Variation in number of general practitioner consultations before hospital referral for cancer: findings from the 2010 National Cancer Patient Experience Survey in England. *Lancet Oncol* 2012;**13**:353–65. http://dx.doi.org/10.1016/S1470-2045(12)70041-4

35. McBride D, Hardoon S, Walters K, Gilmour S, Raine R. Explaining variation in referral from primary to secondary care: cohort study. *BMJ* 2010;**341**:c6267. http://dx.doi.org/10.1136/bmj.c6267

36. Raine R, Wong W, Scholes S, Ashton C, Obichere A , Ambler G. Social variations in access to hospital care for patients with colorectal, breast, and lung cancer between 1999 and 2006: retrospective analysis of hospital episode statistics. *BMJ* 2010;**340**:b5479. http://dx.doi.org/10.1136/bmj.b5479

37. Hippisley-Cox J, Pringle M. Inequalities in access to coronary angiography and revascularisation: the association of deprivation and location of primary care services. *Br J Gen Pract* 2000;**50**:449–54.

38. Cuthbertson SA, Goyder EC, Poole J. Inequalities in breast cancer stage at diagnosis in the Trent region, and implications for the NHS Breast Screening Programme. *J Public Health (Oxf)* 2009;**31**:398–405. http://dx.doi.org/10.1093/pubmed/fdp042

39. Downing A, Prakash K, Gilthorpe MS, Mikeljevic JS, Forman D. Socioeconomic background in relation to stage at diagnosis, treatment and survival in women with breast cancer. *Br J Cancer* 2007;**96**:836–40. http://dx.doi.org/10.1038/sj.bjc.6603622

40. Forrest LF, Adams J, Rubin G, White M. The role of receipt and timeliness of treatment in socioeconomic inequalities in lung cancer survival: population-based, data-linkage study. *Thorax* 2015;**70**:138–45. http://dx.doi.org/10.1136/thoraxjnl-2014-205517

41. Taylor E, Jones F. Lost in translation: exploring therapists' experiences of providing stroke rehabilitation across a language barrier. *Disabil Rehabil* 2014;**36**:2127–35. http://dx.doi.org/10.3109/09638288.2014.892636

42. Hunter DJ. Role of politics in understanding complex, messy health systems: an essay by David J Hunter. *BMJ* 2015;**350**:h1214. http://dx.doi.org/10.1136/bmj.h1214

43. Jusot F, Or Z, Sirven N. Variations in preventive care utilisation in Europe. *Eur J Ageing* 2011;**9**:15–25. http://dx.doi.org/10.1007/s10433-011-0201-9

44. Or Z, Jusot F, Yilmaz E. *Impact of Health Care System on Socioeconomic Inequalities in Doctor Use*. Paris: IRDES; 2008.

45. Cookson R, Asaria M, Ali S, Ferguson B, Fleetcroft R, Goddard M, *et al.* A framework for monitoring NHS equity performance – small area analysis of national administrative data from 2004/5 to 2011/12. *J Epidemiol Community Health* 2015;**69**:A28. http://dx.doi.org/10.1136/jech-2015-206256.43

46. The English Indices of Deprivation 2010. *Department of Communities and Local Government. London, 2011*. URL: www.gov.uk/government/statistics/english-indices-of-deprivation-2010 (accessed 11 April 2016).

47. Ragin CC. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press; 1987.

48. Rihoux B, Lobe B. The Case for Qualitative Comparative Analysis (QCA): Adding Leverage for Thick Cross-Case Comparison. In Byrne D, Ragin CC, editors. *The SAGE Handbook of Case-Based Methods*. London: Sage Publications; 2009. pp. 222–43. http://dx.doi.org/10.4135/9781446249413.n13

49. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage Publications; 1997.

50. Blackman T, Dunstan K. Qualitative comparative analysis and health inequalities: investigating reasons for differential progress with narrowing local gaps in mortality. *J Soc Policy* 2010;**39**:359–73. http://dx.doi.org/10.1017/S0047279409990675

51. Virtanen M, Nyberg ST, Batty GD, Jokela M, Heikkila K, Fransson EI, *et al.* Perceived job insecurity as a risk factor for incident coronary heart disease: systematic review and meta-analysis. *BMJ* 2013;**347**:f4746. http://dx.doi.org/10.1136/bmj.f4746

52. Thomas J, O'Mara-Eves A, Brunton G. Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. *Syst Rev* 2014;**3**:67. http://dx.doi.org/10.1186/2046-4053-3-67

53. Katz JN. Patient preferences and health disparities. *JAMA* 2001;**286**:1506–9. http://dx.doi.org/10.1001/jama.286.12.1506

54. Willems S, De Maesschalck S, Deveugele M, Derese A, De Maeseneer J. Socio-economic status of the patient and doctor-patient communication: does it make a difference? *Patient Educ Couns* 2005;**56**:139–46. http://dx.doi.org/10.1016/j.pec.2004.02.011

55. Karnieli-Miller O, Eisikovits Z. Physician as partner or salesman? Shared decision-making in real-time encounters. *Soc Sci Med* 2009;**69**:1–8. http://dx.doi.org/10.1016/j.socscimed.2009.04.030

56. Hawe P. Lessons from complex interventions to improve health. *Ann Rev Public Health* 2015;**36**:307–23. http://dx.doi.org/10.1146/annurev-publhealth-031912-114421

57. Cutler DM. Equality, efficiency, and market fundamentals: the dynamics of international medical-care reform. *J Econ Lit* 2002;**40**:881–906. http://dx.doi.org/10.1257/jel.40.3.881

58. Tuohy CH. Reform and the politics of hybridization in mature health care states. *J Health Polit Policy Law* 2012;**37**:611–32. http://dx.doi.org/10.1215/03616878-1597448

59. Newhouse JP, Insurance Experiment Group. *Free for All? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press; 1993.

60. Evans RG. Public health insurance: the collective purchase of individual care. *Health Policy* 1987;**7**:115–34. http://dx.doi.org/10.1016/0168-8510(87)90026-1

61. Evans RG, Lomas J, Barer ML, Labelle RJ, Fooks C, Stoddart GL, *et al.* Controlling health expenditures – the Canadian reality. *N Engl J Med* 1989;**320**:571–7. http://dx.doi.org/10.1056/NEJM198903023200906

62. Evans RG, Barer ML, Hertzman C. The 20-year experiment: accounting for, explaining, and evaluating health care cost containment in Canada and the United States. *Ann Rev Public Health* 1991;**12**:481–518. http://dx.doi.org/10.1146/annurev.pu.12.050191.002405

63. Brook RH, Keeler EB, Lohr KN, Newhouse JP, Ware JE, Rogers WH, *et al. The Health Insurance Experiment: A Classic RAND Study Speaks to the Current Health Care Reform Debate*. Santa Monica, CA: RAND Corporation; 2006.

64. Organisation for Economic Co-operation and Development (OECD). *OECD Health Statistics 2015*. Paris: OECD; 2015.

Essay 6 Major system change: a management and organisational research perspective

Simon Turner,[1]* Lucy Goulding,[2]* Jean-Louis Denis,[3]
Ruth McDonald[4] and Naomi J Fulop[1]

[1]Department of Applied Health Research, University College London, London, UK
[2]King's Improvement Science, Centre for Implementation Science, King's College London, London, UK
[3]Canada Research Chair in Governance and Transformation of Health Organisations and Systems, École Nationale d'Administration Publique, Ville de Québec, QC, Canada
[4]Alliance Manchester Business School, University of Manchester, Manchester, UK

*Co-first authors contributed equally to this work

Declared competing interests of authors: Simon Turner and Naomi J Fulop were supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames at Bart's Health NHS Trust. Lucy Goulding is employed by King's Improvement Science. King's Improvement Science is part of the NIHR CLAHRC South London and comprises a specialist team of improvement scientists and senior researchers based at King's College London. Its work is funded by King's Health Partners (Guy's and St Thomas' NHS Foundation Trust, King's College Hospital NHS Foundation Trust, King's College London and South London and Maudsley NHS Foundation Trust), Guy's and St Thomas' Charity, the Maudsley Charity and The Health Foundation. Naomi J Fulop reports grants from NIHR during the conduct of the study, grants from NIHR and grants from The Health Foundation outside the submitted work. Simon Turner reports grants from NIHR outside the submitted work. Lucy Goulding reports grants from King's Health Partners and NIHR outside the submitted work. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Published May 2016
DOI: 10.3310/hsdr04160-85

This essay should be referenced as follows:

Turner S, Goulding L, Denis JL, McDonald R, Fulop NJ. Major system change: a management and organisational research perspective. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 85–104.

List of abbreviations

CLAHRC	Collaboration for Leadership in Applied Health Research and Care
ITS	interrupted time series
RCT	randomised controlled trial
TDF	Theoretical Domains Framework

Abstract

The scale and complexity of major system change in health care (typically involving multiple change processes, organisations and stakeholders) presents particular conceptual and methodological challenges for evaluation by researchers. This essay summarises some current approaches to evaluating major system change from the field of management and organisational research, and discusses conceptual and methodological questions for further developing the field. It argues that multilevel conceptual frameworks and mixed-methods approaches are required to capture the complexity and the heterogeneity of the mechanisms, processes and outcomes of major system change. Future evaluation designs should aim to represent key components of major system change – the context, processes and practices, and outcomes – by looking for ways that quantitative and qualitative methods can enrich one another. Related challenges in ensuring that findings from evaluating major system change are used by decision-makers to inform policy and practice are also discussed.

Scientific summary

The scale and complexity of major system change in health care (typically involving multiple change processes, organisations and stakeholders) presents particular conceptual and methodological challenges for its evaluation by researchers. This essay summarises some current approaches to evaluating major system change from the field of management and organisational research, and discusses conceptual and methodological questions for further developing the field.

Major system change can be seen as a complex intervention; measuring effectiveness is challenging, as contextual factors and processes play a key and often dominant role. In evaluation design, the potential impact of an intervention, and unintended effects, should be captured at different levels of change. Furthermore, rather than seeking to assess 'effectiveness', we advocate exploration of the broader concept of the 'value' of an intervention.

Developing a theoretical framework based on a synthesis of existing theories of organisational change could aid both the design and the evaluation of major system changes. This would help to advance the field through the accumulation of knowledge across projects. However, critical thinking should be employed to ensure that frequently used concepts, including 'leadership' and 'culture', are unpicked and not just taken for granted.

Multilevel conceptual frameworks and mixed-methods approaches are needed to attempt to capture the complexity and the heterogeneity of the mechanisms, processes and outcomes of major system change.

Future evaluation designs should aim to represent key components of major system change – the context, processes and practices, and outcomes – by looking for ways that quantitative and qualitative methods can enrich one another, for example by combining 'what works at what cost' with analysis of 'how' and why' change takes place. To impact on decision-making concerning policy and practice, researchers should

work closely with decision-makers on evaluation design and tailor their findings to different stakeholders, although more focus is needed on overcoming political and structural challenges to collaboration.

Introduction

The complexity and scale of major system change in health care (i.e. typically involving multiple change processes, organisations and stakeholders) presents particular conceptual and methodological challenges for its evaluation by researchers. Drawing on a realist review of literature in this field,[1] we take major system change in health care to involve 'interventions aimed at coordinated, system-wide change affecting multiple organisations and care providers, with the goal of significant improvements in the efficiency of healthcare delivery, the quality of patient care, and population-level patient outcomes'. The authors of the review used the term 'large-system transformation' rather than major system change,[1] but we assume here that the two terms are synonymous. The aim of this essay is to summarise some current approaches to evaluating major system change from the field of management and organisational research, and to discuss conceptual and methodological questions for further developing the field. Related challenges in ensuring that findings from evaluating major system change are used by decision-makers to inform policy and practice are also discussed.

Four talks were given during the plenary on major system change at the London meeting in 2015 described in the Introduction to this volume. Jean-Louis Denis made the case for using process- and practice-based research to evaluate complex interventions in context, in order to maintain the balance with investment in approaches from the clinical sciences that tend to focus on outcome-driven research. Ruth McDonald discussed the use of theory in evaluating major change projects in health systems, suggesting that often taken-for-granted concepts (such as 'leadership' and 'culture') need to be contested and re-evaluated through ongoing critical theory development and in dialogue with empirical research. Naomi J Fulop and Simon Turner described challenges in evaluating major system change at scale using social science theory and methods. They argued that multilevel approaches and mixed methods are necessary for representing the complexity of major system change, but practical challenges for researchers lie in grappling with the scale, significant time and politics often associated with major change programmes. Brian Mittman gave a number of reasons for viewing major system change as a form of complex intervention and the challenges this presents for evaluation, notably that neither interventions nor settings are fixed and their main effects are often weak relative to contextual factors. All four presentations underlined the importance of seeing major system change as complex interventions, that are situated, involve multiple processes of change and operate at multiple levels. There was consensus that multilevel analytical frameworks and mixed-methods approaches are needed to attempt to capture the complexity and the heterogeneity of mechanisms, processes and outcomes of major system change in future evaluations.

Stimulated by the talks and the roundtable discussions that followed at the London meeting, we identified five key themes relating to the evaluation of major system change that we discuss in this essay: (1) type of change and complexity; (2) defining and measuring effectiveness; (3) the role and use of theory in evaluation; (4) the contribution of mixed methods to evaluation; and (5) the use of knowledge from evaluations of major system change to inform policy and practice. Following discussion of each theme, the conclusion summarises the implications, both practical and methodological, for evaluating major system change.

It is important to note that this essay presents a partial view on evaluating major system change, influenced by the particular emphasis of the talks and discussions at the roundtable event, rather than an inclusive, systematic review of the different methodological approaches to evaluating major system change.

Theme 1: type of change and complexity

This theme focuses on the distinctive characteristics of major system change and describes the conceptual and methodological challenges that its complexity and scale present for evaluation.

Characteristics of major system change

Major system change has a number of characteristics that distinguish it from change at a smaller scale, for example change that involves a single organisation or health-care delivery site. Returning to Best *et al.*'s[1] definition, three key characteristics of major system change can be identified. First, major system change often involves the participation of multiple stakeholders, from both within and outside the health-care service. Second, the changes desired are system-wide, meaning that the aim is to produce a collective impact on outcomes across different, often heterogeneous, health-care organisations within a system. Third, major system change involves co-ordinated change over a larger canvas, with mechanisms needed to engage and align stakeholders during the planning and implementation of change, such as leadership, resources or a political mandate from government. Additionally, from an evaluation perspective, major system change can be understood as a complex intervention. It has multiple (sometimes conflicting) goals; it involves change processes at multiple levels; and it takes place within and across heterogeneous settings and often over a significant period of time.

The scale and complexity of major system change creates challenges for its evaluation. It has many of the characteristics of complex interventions. This contrasts with simple interventions that are more likely to have a single fixed component, a stable process, distinct goal, and are applied in relatively homogenous settings, although this is also contested. Major system change involves interventions that change over time, for example they may be adapted on the basis of formative feedback from evaluation; the settings in which they are introduced are not fixed and can be modified; the main effects of the intervention are often weak and contextual factors often dominate; and all three aspects (interventions, settings and effects) can vary over time and space. As we discuss below, evaluation approaches need to be equipped to track this complex set of interactions over time.

Scale and complexity: conceptual implications for evaluations

The distinctive characteristics of major system change when compared with change at a smaller scale, including its relative complexity, imply a range of conceptual and methodological implications for its evaluation by researchers. The first conceptual implication stems from the understanding of major system change as a multilevel process,[2] that is, one involving change processes at the macro level (political, economic and societal context), meso level (organisational) and micro level (sociopsychological behaviour of individuals and groups). When compared with smaller-scale change, major system change is likely to involve significant interaction (and potential tensions) between these different levels. For instance, at the macro level, the external political environment may directly influence change processes, rather than being a mere 'backdrop' to change. For example, a key factor in the implementation of the Scottish Patient Safety Programme was the early involvement of politicians and policy-makers who helped to assemble a national infrastructure to support delivery of the programme.[3] In relation to the reconfiguration of stroke services in major metropolitan areas of England, the 'top-down' implementation of change in London, underpinned by political authority and financial and performance management levers, enabled services to be fully centralised, while a less radical transformation of services took place in Greater Manchester where a more 'bottom-up' (network-based) approach was used.[4]

The second conceptual implication is that, given the need for behaviour to be co-ordinated across a health-care system to achieve major system change, collective organisational structures with political authority, along with the agency of individuals, are likely to be important in the implementation of change. One potential barrier to major system change is the presence of multiple stakeholders' interests associated with the different types of organisations involved. Health-care systems are 'pluralistic settings' in which perceived costs and benefits of change may differ by stakeholder group.[5] Patients, their families and the public are also key stakeholders in the development of major system change programmes.[1] Some evidence suggests that more 'bottom-up' approaches to change may not be appropriate at larger scales of change where multiple stakeholders, with potentially divergent interests, may impede implementation.[6] One implication of this is that complex interventions, and consequently complex change processes, should be thought of as a mix of bottom-up activity with top-down guidance.

A third conceptual implication is that, given the scale of change involved in major system change, collective social processes that transcend organisational boundaries and energise change may play an important role in achieving major system change, such as social movements[7] and collaborative communities of medical professionals.[8] For example, transformation of Denver's health system was aided by political support, including the 'symbolic' role of prominent citizens.[9] The nature of these social processes, and the methods needed to identify and evaluate them, will differ from those associated with the study of face-to-face interactions that are often assumed to influence improvement, for example within clinical micro systems.[10]

Scale and complexity: methodological implications for evaluations

The study of change at a large scale also raises methodological challenges. First, there is the problem of representing change processes across a wide range of settings (e.g. heterogeneous provider and purchasing organisations) and over significant periods of time. Ethnographic case studies, that involve 'thick description'[11] of everyday practices through sustained observation within different sociocultural contexts, appear well suited to generating a detailed understanding of how change processes unfold within a small number of settings over time, but may be more limited in representing the breadth of change processes taking place over a large scale. One way forward is to combine ethnography with methods at other levels, for example wider stakeholder interviews or documentary analysis to capture the macro system context.[12] Context and qualitative methods in health services research are further considered in *Essay 7* in this volume.

Second, major system change programmes may have multiple, often conflicting, goals and may involve multiple components, some of which may be ill-defined by programme leaders, not visible to those carrying out the evaluation, transient in nature or not applied equally across different sites. One way forward is for practitioners and researchers to work closely together in order to build up an understanding of both the intervention's goals and components and its appropriate evaluation. In England, collaboration is being enabled through organisational partnerships such as the National Institute for Health Research Collaborations for Leadership in Applied Health Research and Care (CLAHRCs). However, close collaboration via CLAHRCs has thrown up 'political' challenges, including tensions at times between 'service-centred' and 'research-centred' models of knowledge production.[13]

Theme 2: defining and measuring effectiveness

This theme focuses on the ways in which the effectiveness of major system change might be defined and measured in evaluations. According to the Oxford English dictionary, effectiveness is the degree to which something is successful in producing a desired result. There was much debate about the feasibility of attributing success to 'something' (e.g. an intervention), the measurement of success and identifying outcomes. It was suggested that the characteristics often associated with major system change – its complexity, heterogeneity and instability – renders attempts to evaluate its effectiveness challenging.

Influence of context on an intervention's effectiveness

One of the key challenges in evaluating a complex intervention is the observation that outcomes are often only partially related to the intervention itself; contextual factors/processes play a key and often dominant role. Consequently, it is often impossible to estimate the inherent effectiveness of an intervention deployed as part of a major system change owing to the difficulty of separating the intervention from the context in which it is applied. Even among those who consider that it is possible to decouple the intervention and context, current thinking suggests that it is vital to take into account how contextual factors influence implementation.[14] Even a relatively simple intervention, based on a single fixed component, might have a range of effects depending on the analytical level that is being studied and the co-occurring factors and influences on outcomes. For example, a doctor prescribing an oral antibiotic medicine to a patient might be regarded as a 'simple' intervention, and yet its effectiveness may vary among different people owing to biological (e.g. absorption rate and amount), psychological (motivation), interpersonal (doctor and patient relationship) and wider sociocultural factors.[15] Responses to an intervention or innovation that aims to produce major system change can differ owing to the influence of contextual variables at the system (macro), organisational (meso) and clinical (micro) levels.[16] Taking into account potential change at multiple levels, more recognition and appropriate methods are required to capture the impact of an intervention, as well as its unintended consequences, which may be larger and occur at more levels than was anticipated by programme designers.

Identification of outcomes

The identification of appropriate outcomes or results from major system change was regarded as a key challenge. While often appealing to policy-makers and other stakeholders, it was suggested that single measures of effectiveness (e.g. whether or not an intervention improved a clinical outcome) was unduly narrow and neglected other potential benefits, and unintended consequences, of a given intervention. One suggestion was that a wider concept of 'value' needed to be developed, that goes beyond measuring effectiveness in binary terms (i.e. whether it 'works' or 'does not work'), to capture a broader array of potential benefits and limitations of an intervention. This might include the impact on patient experience, effects on people within organisations and increased understanding of how to manage change. In judging which outcomes to include, it is important to consider the audiences for whom the evaluation is being produced. For instance, a taxpayer of a publicly funded health service such as the NHS may also be interested in 'hard' measures of effectiveness, including cost. In addition to effectiveness and cost, health service managers and researchers are also likely to be interested in what else can be learnt from an intervention programme, for example barriers and enablers to the approach to implementation adopted, and then use this information to adapt the current approach or inform the planning of future programmes.

There was consensus in our discussions that multiple outcomes, that meet the needs of different audiences, should be included in evaluation design and that further work was needed to articulate and agree how wider 'value' beyond effectiveness should be measured. Measuring both effectiveness and value appropriately in different contexts should be informed by greater dialogue between service leaders, researchers and policy-makers. At the service level, researchers can contribute to this dialogue by engaging with how programme designers define 'effectiveness' and whether or not they have a programme theory which might lead them to expect certain impacts.[17] Information on outcomes can also be valuable in acting as a trigger to induce changes in the process and practices deployed by actors and organisations. At the wider policy level, researchers can help to broaden the definition of outcomes or impact beyond effectiveness. For example, an analysis of health-care reform in the UK, Canada and the Netherlands highlights changes in expectations among various publics, in the instruments used to regulate or transform the systems and in the relations of power among concerned publics and key stakeholders.[18]

Measuring outcomes and improving practice

It was suggested that measures of the overall effectiveness of a major change programme should be qualified (e.g. that such outcomes measure results on average) and complemented with finer-grained analyses of experiences of change in a variety of places, among specific stakeholder groups and over different periods of time. Where results deviate from the average, this variation can be used as a source of insight through qualitative research into which contexts and components of an intervention are most effective and why barriers to implementation may emerge in some contexts and not others. However, recognising the interplay between intervention and context raises methodological challenges with measuring the influence of the latter on perceived success, as well as the ways in which the intervention influences the context. For instance, uptake of the Canadian 'Heart Health Kit' (a patient-education resource for preventing cardiovascular disease) among physicians in Alberta was influenced by attributes of the innovation itself as well as contextual and situational factors (e.g. local collegial interaction among potential users).[19]

The embedding of an intervention in a particular context limits the generalisability of the findings to diverse settings. However, if the mechanisms of action associated with an intervention are well understood, this evidence may be transferable to other settings as lessons concerning the mechanisms of action that underpin a given intervention's effectiveness.[20] Here, there is a role for theories in providing external validity of evaluative research of complex interventions. For instance, understanding the interactions between context and interventions implies developing robust theories to examine the relationship between these two.[21] Insights into the effectiveness of an intervention in contexts with particular characteristics could inform future planning: that is whether efforts should be focused on adapting the intervention itself or the underlying context in which it is situated in order to enable improvement.

From the perspective of health-care professionals, relevant questions to improve practice might include understanding how a programme can be adapted and customised in order to increase effectiveness, and how to modify or manage the organisation or setting in order to increase effectiveness. In relation to these questions, it was suggested that the purpose of research is to contribute to the evaluation of a programme's effectiveness by explaining 'how, why, when and where does it work' and, as researchers, address the question of 'how can I make it work better?'. With regard to how the programme operates and why it produces its effects in different contexts, it is important that evaluations take account of potential changes over time in the answers to these 'how' and 'why' questions. Change can be interpreted as either episodic (i.e. radical or exceptional) or continuous (i.e. as an ongoing process of becoming).[22,23] To reflect this, evaluation should seek to analyse how and why an intervention's effects are produced both initially and continuously over time.

Theme 3: the role and use of theory in the evaluation of major system change

This theme outlines the role and value of using theories to inform the evaluation of major system changes. The explicit application of formal theory is needed for both the design and evaluation of major system changes to understand the conditions of context that affect success, to enhance the transferability of learning from changes introduced in one context to another context and therefore to aid accumulation of knowledge across projects.[17] However, the importance of theory in designing, implementing and evaluating interventions arguably remains under-recognised. The ability to informally theorise is often used in day-to-day activities, yet many are alienated by the idea of formally applying theory.[17]

Selecting and applying theories

Different people make sense of phenomena in different ways. A health services researcher will not necessarily view and interpret the mechanisms involved in change and the interaction with the surrounding context in the same way as an anthropologist, geographer, organisational scientist or health policy-maker. Within our occupational silos, we become accustomed to adopting certain familiar theories. We must be aware that adopting a certain theory results in seeing the world through a particular lens: as the adage goes 'when all you have is a hammer, everything looks like a nail'. Our understanding of major system change is shaped by the theories that we use to describe change processes, as individual theories may highlight different parts of the process or encourage similar processes to be interpreted in different ways. Thus, the theoretical standpoint adopted influences the design, conduct and analysis of an evaluation. It is therefore important for evaluators to recognise the importance of selecting appropriate theory and the influence that theory selection will bear on the interpretation of findings. For these reasons, rather than identifying an individual theory to frame a major system change, implementers and evaluators should contemplate adopting multiple relevant theories and ensure that the composition of the team is made up of an appropriate mix of theorists. Having different theoreticians involved increases the number and breadth of questions asked and may help to paint a more diverse and vivid picture of the context in which a major change is delivered.

Deriving insight from theories at different scales

Theories can be broadly categorised as one of three types according to the scale at which they are applied: grand theory, mid-range theory (big theory) and programme theory (small theory). Grand theories are 'formulated at high levels of abstraction' and 'make generalisations that apply across many different domains'.[17] For example, having studied health-care system reforms in the USA, Canada and the UK in the 1990s, Tuohy[24] argues that different patterns of change resulted from the particular logic of each of the systems and that reform was influenced by the distribution of power between different institutions (governments, markets and medical profession) and mechanisms of social control. Mid-range theories 'are intermediate between minor working hypotheses and the all-inclusive speculations comprising a master conceptual scheme'.[17] Examples of mid-range theories include 'diffusion of innovations'[16] and 'normalisation process theory'.[25] Programme theories specify the way in which an intervention is thought to work. They specify the structures (inputs), processes (actions) and outcomes (results) that are anticipated with the links between these providing the theory of change. The influence of behaviours and contextual factors on these components should also be incorporated.[17]

Insight can be derived through dialogue between theories at different scales. For instance, a systematic review of factors affecting innovation adoption in health services categorised these factors according to the theoretical level at which they operate: the sociopolitical climate, system readiness and incentives (grand level); social networks, champions and boundary spanners (mid level); and internal communication, feedback and resources (programme level).[16] Combining theories at different scales may be particularly important in understanding the multilevel influences on major system change and how factors at different levels influence each other (e.g. by mediating or moderating the implementation of a change). As a consequence, multiple levels of analysis are required to assess the links between theory and primary research findings.

Developing theory

Researchers can be criticised for using theory at a level that is too high and too general, taking concepts 'off the shelf' without exploring how applicable these concepts are to the real world. For example, conclusions are often drawn about the importance of 'leadership' and 'culture' without unpacking what it is about these factors that makes them so influential. It was proposed that we need to continue to build theories about what concepts like 'leadership' really mean with regard to major system change. The evaluation of specific programmes, therefore, affords the opportunity to contribute to the refinement or generation of theory at a wider scale.

Using theory in the context of politically charged evaluations

Evaluation of large health system changes places evaluators in a highly political context.[26] Policies are usually the product of political decisions and the recommendations of evaluators are intended to inform policy. Both policy-makers and researchers have an agenda whether or not they are consciously aware of this, highlighting the importance of formally applying theory and making this explicit and accessible to stakeholders. An evaluation may assess whether or not the theory adopted by policy-makers was appropriate. For example, an evaluation of the Commissioning for Quality and Innovation Framework, launched by the Department of Health in England in 2009, sought to refine the theory behind the framework by exploring how it was *envisaged* to work and comparing this with *actual* practice.[27] Policy-makers had based the framework on existing literature and had theorised that financially incentivising clinical teams to set and achieve desired quality targets would be a successful way to encourage behaviour change and thereby drive up quality. In reality, the evaluation team found that the task of selecting quality targets was often undertaken by managers rather than clinical staff, and that staff were concerned that setting quality targets high would put the hospital at financial risk; thus, the focus often shifted from 'high' quality targets to 'achievable' quality targets.[27]

Theme 4: the contribution of different methods to evaluation

Major system changes in health care can be evaluated using a wide range of methods. Within this theme we describe some of these methods, distinguishing between quantitative, qualitative and mixed-methods approaches.

Quantitative approaches to evaluating major system change

Quantitative approaches are used to measure the effectiveness of interventions. Methods adopted include randomised controlled trials (RCTs), natural experiments, interrupted time series (ITS) designs, controlled before-and-after studies and uncontrolled before-and-after studies. The merits of alternative approaches are extensively rehearsed in *Essays 2* and *3* in this volume.

The appeal of a well-designed and well-conducted RCT is that observed differences in outcomes between intervention and control groups can be attributed to the change introduced in the intervention group. Examples of RCTs carried out to evaluate major system change in health care include 'The Health Insurance Experiment', a RCT of the effect of payments for health care on health-care usage and quality of care received conducted in California between 1974 and 1981;[28] 'The Oregon Health Insurance Experiment', an ongoing RCT launched in 2008 that is designed to evaluate the impact of the Medicaid insurance system in the USA on health service use, patient outcomes and economic outcomes;[29–31] the 'Whole System Demonstrators', three large pilots of telehealth launched by the Department of Health in England in 2006 that have been evaluated using a range of methods including a RCT;[32] and the 'Head Start' programme, a RCT of a complex intervention designed to improve the 'school readiness' of children from low-income backgrounds in the USA.[33]

However, there are methodological and implementation challenges in conducting RCTs. Ettelt and Mays[34] discuss these challenges in relation to health policy-making in England. A particular concern regarding results from RCTs is that estimates of effectiveness may not be directly generalisable to the target population of interest if, for example, subgroups of patients/carers are excluded from the trial.[35] The use of observational data is, therefore, advocated to 'assess and strengthen the generalisability of RCT-based estimates of comparative effectiveness'.[35] Furthermore, RCTs are expensive and time-consuming to conduct in comparison with other methods and, in some situations, it is not possible, feasible or ethical to randomise people or organisations to intervention or control conditions.

In 2012 the Medical Research Council published guidance on the use of natural experiments to empirically study population health interventions in situations where randomisation and indeed manipulation of exposure to an intervention is not possible. Natural experiments are therefore defined as 'events, interventions or policies which are not under the control of researchers, but which are amenable to research which uses the variation in exposure that they generate to analyse their impact'.[36]

Interrupted time series designs can be adopted in cases where randomisation is not possible but introduction and rollout of the intervention can be regulated. For example, Yellend *et al.*[37] provide a protocol for an ITS evaluation of a system reform addressing refugee maternal and child health inequalities in Melbourne, Australia. It is preferable for ITS designs to employ concurrent control sites.[38]

Although absence of randomisation is justified in certain circumstances, the use of control groups is essential to estimate whether or not the intervention has a statistically significant impact on outcomes compared with an alternative. For example, Benning *et al.*[39] conducted a RCT of a patient safety programme in the UK known as the 'Safer Patients Initiative'. The trial revealed robust improvements in control sites that were almost as great as the improvements observed in the intervention sites. Had control groups not been in place, it is possible that the improvements in patient safety seen in the intervention group would have been attributed to the intervention and spurious conclusions about effectiveness might have been drawn.

Pronovost and Jha[38] summarise the pitfalls associated with uncontrolled before-and-after studies. Their summary draws on the example of the widely celebrated 'Partnerships for Patients Program' in the USA, which was evaluated using a simple before-and-after design without the use of concurrent controls. They contend that the study design adopted, the absence of valid metrics, the lack of peer review and deficient transparency in reporting make it almost impossible to determine whether or not the programme led to better patient care. Furthermore, before-and-after studies assume that there is a clear definition of what constitutes both 'before' and 'after', although complex interventions are likely to continue to evolve over time and, thus, regular follow-up measurements are recommended.[40]

The appropriateness of different quantitative methods must be considered on a case-by-case basis. Researchers should strive to use the most robust feasible design, adopting control groups and regular measurement, and therefore avoiding uncontrolled before-and-after studies. Those who are evaluating major system changes should be mindful that adopting quantitative methods in isolation may be insufficient. For example, in their discussions of the methodological considerations needed to evaluate the introduction of electronic health records in the English NHS, Takian *et al.*[40] propose the need for an 'interpretive approach' which considers the impact of the national and local context. In addition, Moore *et al.*[41] comment that 'effect sizes do not provide policy-makers with information on how an intervention might be replicated in their specific context, or whether or not trial outcomes will be reproduced', thus highlighting the need to consider qualitative and mixed methods.

Qualitative approaches to evaluating major system change

Experimental and quasi-experimental research designs are appropriate to respond to certain type of research questions and objectives, such as effectiveness, but are not appropriate to respond to research questions that address the 'how' and 'why' questions that modulate the effectiveness of complex interventions in 'natural' settings. Qualitative approaches can help to address these questions by highlighting the contextual factors, and mechanisms of action, that contribute to the effectiveness of an intervention programme, including reasons for results varying across different settings. Process evaluations are useful in accessing the 'black box' of an intervention, that is, understanding the mechanisms through which it produces its effects, and can be used alone or in tandem with other methods (e.g. RCT studies).[15] At the more qualitative end of the evaluation spectrum, case studies are a useful approach for analysing processes of major system change. As described by Yin,[42] case studies involve 'in-depth inquiry into a specific and complex phenomenon (the 'case'), set within its real-world context'. The conduct of case studies involves analysis of the case ('how things are' or an intervention), the context (social, political, financial and so forth), and the interaction between the case and the context and the ways in which they influence one another. These methodological issues are further rehearsed in *Essay 7*.

Rather than describing one method, case studies can be undertaken using a range of methods (including descriptive and theory-driven approaches) and can form a critical component of mixed-methods studies that also aim to explain outcomes.[43] The choice of method depends for the most part on the research questions driving the evaluation. For example, a standalone process evaluation, based on descriptive and theory-driven case studies, was used to conduct a retrospective cross-sectional study of nine health-care mergers in London, as the interest was in how and why the organisational context in which mergers took place, including differences in organisational culture among providers, influenced processes of change.[44] Suggested reporting standards for organisational case studies related to health care have recently been published.[45]

Another theory-driven method is realist evaluation which asks 'what works for whom and under what circumstances?' in an attempt to uncover context–intervention–outcome relationships in change programmes.[46] Realist evaluation recognises that context is complex and always changing. For example, Greenhalgh *et al.*[47] carried out a realist evaluation of the 'modernisation' of stroke, sexual health and kidney services in London. Marchal *et al.*[48] provide a useful overview of studies using realist evaluation in health systems research and propose a need for more methodological guidance on use of the method (e.g. on defining 'mechanisms' and 'context').

Comparative case study research

Although clinical researchers may more easily replicate their observations by studying the effectiveness of a drug on groups of patients, the evaluation of major system change often involves single case studies, making replication of results and assessments of potential generalisability difficult. It is, however, possible to generate cumulative knowledge about the factors that influence major system change through comparative case studies. This involves adopting a structured and theoretically driven approach to summarise, compare and contrast in-depth information derived from two or more studies of major system change.

Comparative research on different major system change programmes can be carried out using studies that were not originally designed with this purpose in mind. For example, Langley *et al.*[49] retrospectively compared cases of large scale health-care transformation in Alberta and Quebec, Canada, in order to explore identity struggles associated with the merger of organisations. Data from a wide range of actors involved in and/or experiencing the change are essential to enable comparisons within and across cases and to generate significant theoretical insights. However, if analysis is planned and undertaken retrospectively, in-depth data of this quality may not be readily available and outputs may not be timely enough to inform practice. On the other hand, prospective data collection is resource intense, requiring considerable time and forward planning between researchers and service leaders to collect as well as risk if proposed changes are not implemented.

Structured comparisons should be made within and across cases and should examine the personnel, process, context and content features of the interventions. For example, Cloutier *et al.*[50] undertook a comparative case study to generate theories about how organisations in Quebec had implemented health-care reform. The study involved a detailed analysis of practice within and between different cases to examine how health-care managers ‘recreate’ reform through conceptual work and testing ideas. The findings suggest that way that people work both dilutes and gives shape to reform at the same time and offers ‘improved understanding of the importance of managerial agency in enacting reform, and the dynamics that lead to slippage in complex reform contexts’.[50]

Hypotheses should be used to structure the comparisons between cases. For example, Øvretveit and Klazinga[51] conducted a mixed-method systematic comparison of factors affecting implementation success of six large-scale quality improvement programmes in the Netherlands. The researchers assessed whether or not there was evidence from their evaluation to support or refute 17 hypotheses about what might predict successful implementation of improvement programmes. The hypotheses were created by an expert team comprising researchers who had worked on the six improvement programmes following a review of the literature. Systematically describing and comparing the different change programmes and their fit (or lack of) with the proposed hypotheses led to the creation of a list of factors thought to be key to the successful implementation of large-scale change. A comparative case study approach has also been used to test hypotheses regarding factors critical to the success of large-scale quality improvement initiatives in Sweden.[52]

Challenges of comparative case study research

Øvretveit and Klazinga[51] discuss four main challenges of comparative case study research into large-scale change, all of which are key considerations for decision-makers when deliberating whether or not apparently successful interventions are replicable elsewhere. Challenges of description arise when limited information is given to describe a change programme, the surrounding context and developments over time. Challenges of attribution arise when the study designs employed (e.g. process research using case studies) make it difficult to say with certainty to what extent a change programme has produced certain outcomes rather than some other factor(s). Challenges of generalisation relate to the extent to which findings from a major system change programme may be generalisable elsewhere. Again, the use of comparative case studies across different programmes can aid assessment of generalisation in different settings. Theories, and research hypotheses that aim to test these, are key to harnessing the cumulative power of doing multiple comparative case studies for the evaluation of complex interventions.

Finally, challenges of use contemplate the utility of research findings for the user. Evaluation designs should strive to be 'useful'; in some situations this may mean that conducting research which gives less-certain answers about whether or not a complex intervention 'works' and more information about the associated processes and context is a better option than conducting a study to answer a single question about effectiveness.

Use of mixed-methods approaches

There is a case for outcomes research, which tends to be more quantitative, and process-based research, which relies more on qualitative methods, to be used in a balanced way, with insights drawn from both approaches and neither dominating. As defined by Langley *et al.*,[53] process studies 'address questions about how and why things emerge, develop, grow, or terminate over time', which differs from quantitative studies that tackle 'variance questions dealing with covariation among dependent and independent variables'. Evaluations involving mixed methods are one way of bringing together quantitative and qualitative analyses of major system change. For example, mixed methods have been used to evaluate reconfiguration of acute stroke services across two large metropolitan areas in England (London and Greater Manchester), combining quantitative analysis of the impact of change on patient outcomes and cost using a controlled before-and-after study with a process evaluation of 'how' and 'why' different approaches to the planning and implementation of change were adopted in each area.[54] Additionally, mixed-methods approaches have been used to evaluate the Advancing Quality pay-for-performance programme in North-West England[55] and a large-scale transformational change programme in the North East.[56]

Theme 5: using knowledge from evaluation to inform policy and practice

This theme focuses on the challenges that arise in ensuring that findings from major system change evaluations are used by decision-makers to inform policy and practice. The creation of knowledge about how major system change programmes can best be delivered and evaluated and the informed assessment of the potential transferability of successful major system change programmes to other settings are central to the impact of management and organisational research in health care.

Accumulation of knowledge: theoretical and methodological considerations

The third theme in this essay demonstrated the importance of employing theory. Theory allows inferences about social and organisational process to be extrapolated beyond a surface description of individual cases, thus playing a vital role in the assessment of potential transferability of findings from a given context to another context and aiding the accumulation of knowledge across studies.

The sections on qualitative and mixed methods within the fourth theme of this essay detail the contribution of process- and practice-based research to evaluate major system change in health care. Case studies are often employed to describe how and why change emerges (and potentially disappears) over time. A particular challenge of case study designs of this type is moving beyond the production of idiosyncratic descriptive accounts of change towards a deeper understanding of the underlying generative mechanisms that interact to create change. It is essential to explore how to gain generalisable insights that have value in terms of transferability beyond the original context in which major system change was deployed in order that others might improve their own systems with increased ease and efficiency. Comparative case study research can be a useful method in this regard, as described in the fourth theme of this essay.

Creating actionable findings

When discussing how to build actionable knowledge in the field of major system change in health care, participants suggested that a potential way forward would be to create a theoretical framework based on a synthesis of existing theories of organisational change. Such a framework would resemble the

Theoretical Domains Framework (TDF) which was developed by Michie *et al.*[57–60] to make behaviour change theories accessible for implementation researchers. The TDF contains constructs from 33 different theories that explain behaviour change in individuals. It enables researchers to use theory to target the behaviour change of patients or health-care professionals by providing operational guidance on the development of complex interventions that are designed to reduce the gap between clinical practice and the evidence base.[61] The benefits of creating a framework containing constructs from organisational change theories would be twofold: the framework would provide a theoretically informed 'road map' for those implementing major system changes and would essentially create a feedback loop, allowing researchers to use findings from one study to inform another.

Communicating findings

The likelihood that findings from major system change evaluations are used by decision-makers to inform policy and practice is linked to the ability of evaluators to communicate their findings. For maximum impact, researchers must tailor the dissemination of their findings to different stakeholders. Furthermore, the dissemination piece must describe and debate issues relating to the implementation of the major system change, rather than solely focusing on headline-grabbing statistics relating to outcomes, in order that decision-makers and other stakeholders can assess the potential for transferability to other settings. However, researchers can, arguably, be poor at communicating, marketing and 'selling' their knowledge, and could perhaps learn lessons from the large health-care consultancies in this regard. The way in which academic careers are structured, with a particular focus on peer-reviewed journal articles, can be a disincentive to invest time in dissemination via other channels of communication, including those channels that are likely to be accessed by decision-makers. Further investment in science communication could help to correct this imbalance and therefore drive an increase in evidence-based approaches to major system change.

The role of researchers: collaborators or evaluators?

Decision-makers do not solely rely on research evidence when making decisions about the major changes required to improve health care and how these changes should be implemented. Policy documents, data collected by organisations, journalistic accounts and anecdotal accounts are important sources of information that are likely to influence decisions regarding major system change. Rigorous evaluation that applies research methods is time-consuming and costly relative to data gathering using these alternative forms of evidence. One school of thought suggests that it is, therefore, necessary for researchers to work more dynamically and in closer partnership with other stakeholders in order to inform and strengthen the implementation and evaluation of ongoing change. However, there can be political challenges associated with working in this way, as outlined in this essay. Furthermore, another school of thought advocates that researchers should maintain a critical distance in order to evaluate objectively.

The predominant view expressed at the roundtable event was that researchers should play one of two roles: they should either 'sit at the table' with decision-makers and help them to plan the interventions and their implementation (but not conduct the evaluation of this) or independently conduct an evaluation. Thus, it can be proposed that there is a need for separate implementation and evaluation teams in order to maintain transparency of the evaluation methodology. Nevertheless, where introduction of a major system change and evaluation activities take place contemporaneously, researchers should strive to produce evaluation findings in a timely manner in order to guide the ongoing implementation of the change to enhance relevance and impact of the evaluation. Researchers have produced guidance to help decision-makers to decide whether or not and how to introduce change.[62] Considerations of implementation are further addressed in *Essay 8*.

Expectations of researchers versus expectations of decision makers

Researchers are usually commissioned to conduct evaluations of major system changes in health care by decision-makers who are the driving force behind the change. Challenges may arise for researchers when they are asked to evaluate a change that is not readily 'evaluable' or when there is not clear consensus between what decision-makers want and what researchers are able to offer. For example, decision-makers may hope that researchers are able to evaluate the effectiveness of a complex major system change at a whole-system level, yet it is not always possible to deliver this. There then begins a negotiation between decision-makers and researchers to determine and agree what is possible and meaningful for both parties.

It is also important to establish a shared understanding of the way in which evaluation findings will be used from the outset. For example, different stakeholders may have different understandings of the term 'pilot study'.[63] In the eyes of the decision-maker, a pilot study may strongly signify the intended direction of travel. The evaluation is intended to shape but not dramatically alter this direction of travel: it is the oar that steers the boat. On the other hand, a researcher may understand the words 'pilot study' to mean a test to determine whether or not the intervention should be continued. They may expect an intervention to be terminated if the findings of the pilot study suggest that the change is not effective or is having unintended consequences. In reality, this may not be the case. For example, an urgent care telephone triage service known as 'NHS 111' was introduced as a pilot in four geographically defined areas in England in 2010. The aim of the telephone service was to 'improve access to urgent care, increase efficiency by directing people to the "right place first time" including self-care advice, increase satisfaction with urgent care and the NHS generally, and in the longer term reduce unnecessary calls to the 999 emergency ambulance service and so begin to rectify concerns about the inappropriate use of emergency services'[64] [quote reproduced as this is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/]. An evaluation of the pilot demonstrated that the introduction of the NHS 111 telephone service actually increased ambulance use in pilot sites in comparison with control sites.[64] Despite the negative findings uncovered in the pilot evaluation, the NHS 111 telephone service was subsequently rolled out nationwide.

Ettelt *et al.*[65] provide further reflections on the tensions that can arise as a result of the different perspectives held by researchers and decision-makers.

Conclusion

There was broad consensus from the meeting that major system change is a complex and unstable process, which operates at multiple levels, and is context dependent and thus varies by place and over time. The complexity, heterogeneity and instability of major system change also presents challenges for defining and measuring its effectiveness, although a range of approaches are being used from quantitative measures of outcomes through to broader measures of the 'value' of change used in qualitative and process-based studies. The use of different methods and perspectives to study major system change, and their combination in the design of evaluations (e.g. mixed-methods approaches), is important in order to represent different aspects of the complexity of major system change and to evaluate its effectiveness, which should be broadly defined in terms of potential impacts.

Theory allows inferences about social and organisational process to be extrapolated beyond a surface description of individual cases, thus playing a vital role in the assessment of potential transferability of findings from a given context to another context and aiding the accumulation of knowledge across studies. The creation of a theoretical framework based on a synthesis of existing theories of organisational change could aid both the design and the evaluation of major system changes and help to advance the field through accumulation of knowledge across projects. However, there is a tension between, on the one hand, the need to accumulate a stable body of knowledge (e.g. on implementation) that could be used to inform policy and practice and, on the other, the maintenance of critical thinking in relation to existing theory to ensure that concepts that have achieved widespread acceptance are not just taken for granted, but remain provisional and open to challenge.

There are implications for furthering the development of both process-based and outcome-based studies of major system change, as well as identifying and pursuing novel ways of bringing the two approaches together. In particular, future evaluation designs should aim to represent and capture key components of the dynamics of major system change – the context, process and practices, and outcomes – as understanding of these elements is currently held back by the tendency to specialise in either qualitative or quantitative methods of research, rather than look for common ground where they can enrich one another.

Acknowledgements

Contributions of authors

Simon Turner (Senior Research Associate, Health Services Research) and **Lucy Goulding** (Post-Doctoral Researcher, Health Services Research) wrote the first draft of the essay, under the guidance of **Naomi J Fulop** (Professor, Health Care Organisation and Management).

Jean-Louis Denis (Professor, Health Systems Research) and **Ruth McDonald** (Professor, Health Science Research and Policy) commented on the draft and provided additional material.

All authors approved the final version of the essay.

References

1. Best A, Greenhalgh T, Lewis S, Saul JE, Carroll S, Bitz J. Large-system transformation in health care: a realist review. *Milbank Q* 2012;**90**:421–56. http://dx.doi.org/10.1111/j.1468-0009.2012.00670.x
2. Rousseau DM. Reinforcing the micro/macro bridge: organizational thinking and pluralistic vehicles. *J Manage* 2011;**37**:429–42. http://dx.doi.org/10.1177/0149206310372414
3. Haraden C, Leitch J. Scotland's successful national approach to improving patient safety in acute care. *Health Aff (Millwood)* 2011;**30**:755–63. http://dx.doi.org/10.1377/hlthaff.2011.0144
4. Turner S, Ramsay A, Perry C, Boaden R, McKevitt C, Morris S, *et al.* Lessons for major system change: centralization of stroke services in two metropolitan areas of England [published online ahead of print 24 January 2016]. *J Health Serv Res Policy* 2016. http://dx.doi.org/10.1177/1355819615626189
5. Langley A, Denis J-L. Beyond evidence: the micropolitics of improvement. *BMJ Qual Saf* 2011;**20**:i43–6. http://dx.doi.org/10.1136/bmjqs.2010.046482
6. Conrad DA, Grembowski D, Hernandez SE, Lau B, Marcus-Smith M. Emerging lessons from regional and state innovation in value based payment reform: balancing collaboration and disruptive innovation. *Milbank Q* 2014;**92**:568–623. http://dx.doi.org/10.1111/1468-0009.12078
7. Waring J. *A Movement for Improvement? A Qualitative Study on the Use of Social Movement Strategies in the Implementation of a Quality Improvement Intervention*. Presentation at Health Services Research Network Symposium, Nottingham Conference Centre, Nottingham, UK, 1–2 July 2015.
8. Adler PS, Kwon SW, Heckscher C. Perspective-professional work: the emergence of collaborative community. *Organ Sci* 2008;**19**:359–76. http://dx.doi.org/10.1287/orsc.1070.0293
9. Harrison MI, Kimani J. Building capacity for a transformation initiative: system redesign at Denver Health. *Health Care Manage Rev* 2009;**34**:42–53. http://dx.doi.org/10.1097/01.HMR.0000342979.91931.d9

10. Barach P, Johnson JK. Understanding the complexity of redesigning care around the clinical microsystem. *Qual Saf Health Care* 2006;**15**(Suppl. 1):10–16. http://dx.doi.org/10.1136/qshc.2005.015859

11. Geertz C. Thick Description: Toward an Interpretive Theory of Culture. In Lincoln Y, Denzin N, editors. *Turning Points in Qualitative Research: Tying Knots in a Handkerchief*. Oxford: Altamera Press; 2003. pp.143–68.

12. Robert GB, Anderson JE, Burnett SJ, Aase K, Andersson-Gare B, Bal R, *et al.* A longitudinal, multi-level comparative study of quality and safety in European hospitals: the QUASER study protocol. *BMC Health Serv Res* 2011;**11**:285. http://dx.doi.org/10.1186/1472-6963-11-285

13. Currie G, Lockett A, El Enany N. From what we know to what we do: lessons learned from the translational CLAHRC initiative in England. *J Health Serv Res Policy* 2013;**18**(Suppl. 3):27–39. http://dx.doi.org/10.1177/1355819613500484

14. Eccles MP, Armstrong D, Baker R, Cleary K, Davies H, Davies S, *et al.* An implementation research agenda. *Implement Sci* 2009;**4**:1–7. http://dx.doi.org/10.1186/1748-5908-4-18

15. Richards DA. The Complex Intervention Framework. In Richards DA, Hallberg, IR, editors. *Complex Interventions in Health: An Overview of Research Methods*. London: Routledge; 2015. pp. 1–15.

16. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. *Milbank Q* 2004;**82**:581–629. http://dx.doi.org/10.1111/j.0887-378X.2004.00325.x

17. Davidoff F, Dixon-Woods M, Leviton L, Michie S. Demystifying theory and its use in improvement. *BMJ Qual Saf* 2015;**24**:228–38. http://dx.doi.org/10.1136/bmjqs-2014-003627

18. Tuohy CH. Reform and the politics of hybridization in mature health care states. *J Health Polit Policy Law* 2012;**37**:611–32. http://dx.doi.org/10.1215/03616878-1597448

19. Scott SD, Plotnikoff RC, Karunamuni N, Bize R, Rodgers W. Factors influencing the adoption of an innovation: an examination of the uptake of the Canadian Heart Health Kit (HHK). *Implement Sci* 2008;**3**:41. http://dx.doi.org/10.1186/1748-5908-3-41

20. Dixon-Woods M, Bosk C, Aveling EL, Goeschel CA, Pronovost PJ. Explaining Michigan: developing an ex post theory of a quality improvement program. *Milbank Q* 2011;**89**:167–205. http://dx.doi.org/10.1111/j.1468-0009.2011.00625.x

21. Fulop N, Robert G. *Context for Successful Improvement: Evidence Review*. London: The Health Foundation; 2015.

22. Tsoukas H, Chia R On organizational becoming: rethinking organizational change. *Organ Sci* 2002;**13**:567–82. http://dx.doi.org/10.1287/orsc.13.5.567.7810

23. Langley A, Denis J-L. [Les dimensions négligées du changement organisationnel.] *Télescope* 2008;**14**:13–32.

24. Tuohy C. *Accidental Logics: The Dynamics of Policy Change in the United States, Britain and Canada*. Oxford: Oxford University Press; 1999.

25. May C. Towards a general theory of implementation. *Implement Sci* 2013;**8**:18. http://dx.doi.org/10.1186/1748-5908-8-18

26. Weiss C. *Evaluation: Methods for Studying Programs and Policies*. 2nd edn. Upper Saddle River, NJ: Prentice Hall; 1998.

27. McDonald R, Kristensen SR, Zaidi S, Sutton M, Todd S, Konteh F, *et al. Evaluation of the Commissioning for Quality and Innovation Framework Final Report*. Manchester: University of Manchester; 2013. URL: http://hrep.lshtm.ac.uk/publications/CQUIN_Evaluation_Final_Feb2013–1.pdf (accessed August 2015).

28. Brook RH, Keeler EB, Lohr KN, Newhouse JP, Ware JE, Rogers WH, *et al. The Health Insurance Experiment: A Classic RAND Study Speaks to the Current Health Care Reform Debate*. Santa Monica, CA: RAND Corporation; 2006. URL: www.rand.org/pubs/research_briefs/RB9174 (accessed August 2015).

29. Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, *et al.* The Oregon Health Insurance Experiment: evidence from the first year. *Q J Econ* 2012;**127**:1057–106. http://dx.doi.org/10.1093/qje/qjs020

30. Taubman S, Allen H, Wright B, Baicker K, Finkelstein A. Medicaid increases emergency department use: evidence from Oregon’s Health Insurance Experiment. *Science* 2014;**343**:263–8. http://dx.doi.org/10.1126/science.1246183

31. Baicker K, Finkelstein A, Song J, Taubman S. *The Impact of Medicaid on Labor Force Activity and Program Participation: Evidence from the Oregon Health Insurance Experiment.* NBER working paper 19547. Cambridge, MA: National Bureau of Economic Research; 2013.

32. Steventon A, Bardsley M, Billings J, Dixon J, Doll H, Hirani S, *et al.* Effect of telehealth on use of secondary care and mortality: findings from the Whole System Demonstrator cluster randomised trial. *BMJ* 2012;**344**:e3874. http://dx.doi.org/10.1136/bmj.e3874

33. US Department of Health and Human Services. *Head Start Impact Study Final Report*. Washington, DC: US Department of Health and Human Services; 2010.

34. Ettelt S, Mays N. RCTs – how compatible are they with contemporary health policy-making? *Br J Health Manag* 2015;**21**:379–82. http://dx.doi.org/10.12968/bjhc.2015.21.8.379

35. Steventon A, Grieve R, Bardsley M. An approach to assess generalizability in comparative effectiveness research: a case study of the whole systems demonstrator cluster randomized trial comparing telehealth with usual care for patients with chronic health conditions. *Med Decis Making* 2015;**35**:1023–36. http://dx.doi.org/10.1177/0272989X15585131

36. Medical Research Council. *Using Natural Experiments to Evaluate Population Health Interventions: Guidance for Producers and Users Of Evidence*. MRC; 2012. URL: www.mrc.ac.uk/naturalexperimentsguidance (accessed August 2015).

37. Yelland J, Riggs E, Szwarc J, Casey S, Dawson W, Vanpraag D, *et al.* Bridging the gap: using an interrupted time series design to evaluate systems reform addressing refugee maternal and child health inequalities. *Implement Sci* 2015;**10**:62. http://dx.doi.org/10.1186/s13012-015-0251-z

38. Pronovost P, Jha AK. Did hospital engagement networks actually improve care? *N Eng J Med* 2014;**371**:691–3. http://dx.doi.org/10.1056/NEJMp1405800

39. Benning A, Dixon-Woods M, Nwulu U, Ghaleb M, Dawson J, Barber N, *et al.* Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ* 2011;**342**:d199. http://dx.doi.org/10.1136/bmj.d199

40. Takian A, Dimitra P, Cornford T, Sheikh A, Barber N. Building a house on shifting sand: methodological considerations when evaluating the implementation and adoption of national electronic health record systems. *BMC Health Serv Res* 2012;**12**:105. http://dx.doi.org/10.1186/1472-6963-12-105

41. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, *et al.* Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 2015;**350**:h1258. http://dx.doi.org/10.1136/bmj.h1258

42. Yin RK. Validity and generalization in future case study evaluations. *Evaluation* 2013;**19**:321–32. http://dx.doi.org/10.1177/1356389013497081

43. Yin RK. *Case Study Research Design and Methods*. 4th edn. London: Sage; 2009.

44. Fulop N, Protopsaltis G, King A, Allen P, Hutchings A, Normand C. Changing organisations: a study of the context and processes of mergers of health care providers in England. *Soc Sci Med* 2005;**60**:119–30. http://dx.doi.org/10.1016/j.socscimed.2004.04.017

45. Rodgers M, Thomas S, Harden M, Parker G, Street A, Eastwood A. Developing a methodological framework for organisational case studies: a rapid review and consensus development process. *Health Serv Deliv Res* 2016;**4**(1).

46. Pawson R, Tilley N. *Realistic Evaluation*, London: Sage; 1997.

47. Greenhalgh T, Humphrey C, Hughes J, Macfarlane F, Butler C, Pawson R. How do you modernize a health service? A realist evaluation of whole-scale transformation in London. *Milbank Q* 2009;**87**:391–416. http://dx.doi.org/10.1111/j.1468-0009.2009.00562.x

48. Marchal B, van Belle S, van Olmen J, Hoerée T, Kegels G. Is realist evaluation keeping its promise? A review of published empirical studies in the field of health systems research. *Evaluation* 2012;**18**:192–212. http://dx.doi.org/10.1177/1356389012442444

49. Langley A, Golden-Biddle K, Reay T, Denis J-L, Hébert Y, Lamothe L, *et al.* Identity struggles in merging organizations: renegotiating the sameness–difference dialectic. *J Appl Behav Sci* 2012;**48**:135–67. http://dx.doi.org/10.1177/0021886312438857

50. Cloutier C, Denis J-L, Langley A, Lamothe L. Agency at the managerial interface: public sector reform as institutional work [published online ahead of print 1 June 2015]. *J Public Adm Res Theory* 2015.

51. Øvretveit J, Klazinga N. Learning from large-scale quality improvement through comparisons. *Int J Qual Health Care* 2012;**24**:463–9. http://dx.doi.org/10.1093/intqhc/mzs046

52. Øvretveit J, Andreen-Sachs M, Carlsson J, Gustafsson H, Hansson J, Keller C, *et al.* Implementing organisation and management innovations in Swedish healthcare: lessons from a comparison of 12 cases. *J Health Organ Manag* 2012;**26**:237–57. http://dx.doi.org/10.1108/14777261211230790

53. Langley A, Smallman C, Tsoukas H, Van de Ven AH. Process studies of change in organization and management: unveiling temporality, activity, and flow. *Acad Manage J* 2013;**56**:1–13. http://dx.doi.org/10.5465/amj.2013.4001

54. Fulop N, Boaden R, Hunter R, McKevitt C, Morris S, Pursani N, *et al.* Innovations in major system reconfiguration in England: a study of the effectiveness, acceptability and processes of implementation of two models of stroke care. *Implement Sci* 2013;**8**:19. http://dx.doi.org/10.1186/1748-5908-8-5

55. McDonald R, Boaden R, Roland M, Kristensen SR, Meacock R, Lau Y-S, *et al.* A qualitative and quantitative evaluation of the Advancing Quality pay-for-performance programme in the NHS North West. *Health Serv Deliv Res* 2015;**3**(23).

56. Hunter DJ, Erskine J, Hicks C, McGovern T, Small A, Lugsden E, *et al.* A mixed-methods evaluation of transformational change in NHS North East. *Health Serv Deliv Res* 2014;**2**(47).

57. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A. 'Psychological Theory' Group. Making psychological theory useful for implementing evidence based practice: a consensus approach. *Qual Saf Health Care* 2005;**14**:26–33. http://dx.doi.org/10.1136/qshc.2004.011155

58. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioral determinants to behavior change techniques. *Applied Psychol* 2008;**57**:660–80. http://dx.doi.org/10.1111/j.1464-0597.2008.00341.x

59. Michie S, Fixsen D, Grimshaw JM, Eccles MP. Specifying and reporting complex behaviour change interventions: the need for a scientific method. *Implement Sci* 2009;**4**:40. http://dx.doi.org/10.1186/1748-5908-4-40

60. Michie S, Atkins L, West R. *The Behaviour Change Wheel. A Guide to Developing Interventions*. London: Silverback Publishing; 2014.

61. French SD, Green SE, O'Connor DA, McKenzie JE, Francis JJ, Michie S, *et al.* Developing theory-informed behaviour change interventions to implement evidence into practice: a systematic approach using the Theoretical Domains Framework. *Implement Sci* 2012;**7**:38. http://dx.doi.org/10.1186/1748-5908-7-38

62. Brach C, Lenfestey N, Roussel A, Amoozegar J, Sorensen A. *Will It Work Here? A Decision maker's Guide to Adopting Innovations*. Rockville, MD: Agency for Healthcare Research and Quality; 2008.

63. Ettelt S, Mays N, Allen P. The multiple purposes of policy piloting and their consequences: three examples from national health and social care policy in England. *J Soc Policy* 2015;**44**:319–33. http://dx.doi.org/10.1017/S0047279414000865

64. Turner J, O'Cathain A, Knowles E, Nicholl J. Impact of the urgent care telephone service NHS 111 pilot sites: a controlled before and after study. *BMJ Open* 2013,**3**:e003451. http://dx.doi.org/10.1136/bmjopen-2013-003451

65. Ettelt S, Mays N, Allen P. Policy experiments: investigating effectiveness or confirming direction? *Evaluation* 2005;**21**:292–307. http://dx.doi.org/10.1177/1356389015590737

Essay 7 Contextual issues and qualitative research

Emma Howarth,[1] Kelly Devers,[2] Graham Moore,[3]
Alicia O'Cathain[4] and Mary Dixon-Woods[5]

[1]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) East of England, University of Cambridge, Cambridge, UK
[2]Health Policy Centre, Urban Institute, Washington, DC, USA
[3]School of Social Sciences, Cardiff University, Cardiff, UK
[4]School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[5]Department of Health Sciences, University of Leicester, Leicester, UK

Declared competing interests of authors: Kelly Devers served as the Urban Institute Project Director for the Children's Health Insurance Program Reauthorization Act Quality Demonstration Evaluation briefly described in the manuscript. She also served as the qualitative, cross-state and Programme Assistance team lead on the Centre for Medicare and Medicaid Innovation Multipayer Advanced Primary Care Practice Demonstration Evaluation briefly described in the manuscript. However, her observations in this essay are solely her own and do not reflect the opinions or perspectives of any federal agency or the Urban Institute.

Published May 2016
DOI: 10.3310/hsdr04160-105

This essay should be referenced as follows:

Howarth E, Devers K, Moore G, O'Cathain A, Dixon-Woods M. Contextual issues and qualitative research. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 105–20.

List of tables

List of figures

List of abbreviations

CONSORT	Consolidated Standards of Reporting Trials
ICU	intensive care unit
MRC	Medical Research Council
NSQIP	National Surgical Quality Improvement Program
QIC	quality improvement collaborative
RCT	randomised controlled trial

Abstract

The need to understand the contextual influences that subvert, neutralise, intensify or otherwise influence interventions is well recognised, yet the role that context plays in the design and evaluation of health and social care interventions remains understudied. Concerted effort from the research community – researchers, funders and publishers – is needed to address this problem. Mixed-methods studies are likely to be critical. Researchers should expend greater effort earlier in the development and evaluation process to optimise interventions, identify causal mechanisms and characterise contextual influences. The development of quantitative measures to explore structural features of context should be a future focus, as should better reporting of interventions and the development of guidance to support explicit descriptions of context. Encouraging the conduct of mixed-method process evaluations alongside trials and other evaluations will require new norms and expectations of research and associated funding commitments, attention to collaborative relationships and changes in publishing practices.

Scientific summary

The need to understand the contextual influences that subvert, neutralise, intensify or otherwise influence interventions is increasingly well recognised, yet the role that context plays in the design and evaluation of health and social care interventions has remained understudied. Concerted effort is needed to address the conceptual, methodological and practical challenges associated with conducting contextually sensitive research.

Several key issues present barriers to progress. First, poor-quality reporting of interventions (and programmes more broadly) hampers understanding of what an intervention comprises, the mechanisms through which it works and the likely influences on implementation, including those of context. Second, the need to understand and give accounts of context is continually challenged by the problem of conceptualising what ‘context’ means in practice. Third, inadequate time is given over in the early stages of research to theorising about context.

Effort expended early in the development and evaluation process is needed to optimise the theoretical basis of interventions and to characterise the systems into which they will be introduced. Quantitative measures to explore structural features of context are essential; more effort in helping to specify these measures and support data collection will be an important focus for the future. Qualitative methods are particularly suited to securing understanding of the theoretical basis of interventions, exploring how, where, when and by whom an is intervention delivered and received, identifying effects and how they are achieved, and characterising how interventions are adapted in response to contextual influences and how contexts themselves are modified by interventions. For the future, researchers, funders and publishers should be encouraged to think in terms of mixed-methods studies where qualitative and quantitative methods are seen as complementary methods. Better reporting of interventions and the development of guidance to support explicit descriptions of context are also needed.

Realisation of these ideas will require funder commitments to adequately resource both developmental work and process evaluations that run alongside trials and evaluation designs, more realistic expectations about the time frame for developing an intervention, and journals that are willing to publish the results of this kind of work. The institutional features of the research landscape – including funding structures, publishing models, and the norms, values and practices of the researcher community and the users of evidence – also need critical scrutiny to ensure that they support contextually sensitive research.

Introduction

Medical Research Council (MRC) guidance on developing and evaluating complex interventions recommends that researchers should 'first develop the intervention to the point where it can reasonably be expected to have a worthwhile effect'.[1] But health and social care systems continue to be challenged by the problem that initially promising interventions often prove difficult to replicate and scale: what works in one setting may not be as effective, and may sometimes even be harmful, when implemented elsewhere.[1] Efforts to understand the contextual influences that may subvert, neutralise, intensify or otherwise influence interventions are, therefore, critical. Yet current practices of research and evaluation are pervaded by a tendency to defer thinking about 'contextual contingencies' until too late.[2]

Participants at the London meeting in 2015, which is described in the *Introduction* to this volume, suggested that the sparse attention given to context in the design and evaluation of health and social care interventions[3] is explained not by the perception that context is unimportant, but by the formidable conceptual and practical challenges associated with defining, measuring and articulating the role of contextual factors and influences. Even a consensual definition of context remains elusive. The need for serious attention to questions of context is now, however, increasingly well recognised across a range of disciplines,[4,5] to the extent that participants emphasised the need for guidance on how to structure thinking about context and to select appropriate methods for its study.

In this essay, we propose that moving the field forward will require more sophisticated characterisation of interventions and their contexts of implementation. We emphasise the benefits of quantitative measures, but highlight the value of qualitative methods and of combining qualitative and quantitative methods when addressing questions of context.

Understanding and reporting interventions

One major challenge in studying context and contextual influences is the persistent problem of poor-quality reporting of interventions and programmes. Good accounts of what a programme/intervention is, how it works and how it is influenced by and exerts influence on the surrounding context are rare.[6] Although the Consolidated Standards of Reporting Trials (CONSORT) requires that 'the interventions for each group with sufficient details to allow replication, including how and when they were actually administered' (item 5)[7] be reported, only a minority of non-pharmacological interventions are adequately described.[8,9] Without a complete and explicit description of an intervention,[11] clinicians, patients, researchers and other decision-makers are left unclear about what an intervention comprises and the likely influences on implementation, including those of context.[11] This essay therefore expands on considerations regarding trial design and evaluation discussed in *Essay 2* in this volume.

Publication guidelines now increasingly support the view that a complete description means that that the theoretical basis of interventions must be reported alongside the components of interventions,[11] not least because this allows an assessment of how far the intervention is theoretically well founded. It may also help in explaining the findings of studies and in optimising the design and implementation of interventions in the light of those findings. For instance, the American College of Surgeons National Surgical Quality Improvement Program (NSQIP), a large-scale audit and feedback exercise that provides feedback to hospitals on their risk-adjusted outcomes (e.g. mortality, specific complications and length of stay), was found in a quasi-experimental study to have no additional effect on the outcomes of participating hospitals relative to matched controls after taking into account pre-existing secular trends.[12] A possible explanation for this disappointing result is that the audit and feedback mechanism used in NSQIP is not fully consistent with the findings of the significant body of research and theory on features of feedback effectiveness, including those relating to frequency of feedback, selection of solutions and setting of goals and action plans.[13–15]

Use of theory can also help in explaining why interventions may fare differently in different contexts, as shown by studies of the quality improvement collaborative (QIC) model. The model requires organisations to come together over a defined period of time to engage in collaborative learning and exchange of insights to achieve improvement goals.[16,17] Although QIC models have proliferated widely,[18] understanding of whether or not they work – and, if so, how, when and where they work best – has remained underdeveloped.[17] Here, again, theories are important. An evaluation of a QIC model to improve stroke care in the north-west of England[16] using a randomised design to assess performance on nine aspects of stroke care found that hospitals participating in the QIC model showed only modest improvements on selected aspects of care compared with non-participants. A qualitative study to explore the views and experiences of those involved in designing and implementing the QIC model found that it conferred some benefits but also introduced significant tensions – many of them possible to predict using theory on collective action.[19] For example, participation required substantial effort on the part of participating organisations, introducing the risk of 'free-riding' where some organisations gave less than they received. Importantly, this qualitative study also identified the influences on organisations' commitment to the collaborative ethos by characterising features of both the inner context of individual hospitals (e.g. baseline performance) and an outer context of competition over the provision of stroke services.

When descriptions of interventions and accounts of how they work are inadequate, attempts to replicate successful programmes risk disappointment that may be blamed on context, when the true culprit is that only the superficial outer appearance of the intervention has been reproduced and not the mechanisms (or set of mechanisms) that produced the outcomes in the first instance: a phenomenon known as 'cargo cult science'.[10] This was exemplified when the success of a complex intervention that reduced central venous catheter bloodstream infections across a large number of intensive care units (ICUs) in the state of Michigan[20] was simplistically credited to the implementation of a basic checklist.[21] A post hoc analysis to discern the mechanisms through which the programme worked revealed a complex web of interacting social processes, including the technical and cultural shifts that ultimately increased patient safety.[10] A qualitative study of the attempt to replicate the programme in England suggested that the mechanisms that underpinned the programme's success in Michigan were generally not activated in the English version, even though its components were similar.[22]

These findings, like those of other studies, add weight to the recent suggestion that the fidelity of an intervention may reside in the mechanisms of action rather than in specific programme components, such that individual components of an intervention may be varied across contexts without comprising the effects of the programme.[23,24] This requires an understanding of the role and meaning of each intervention component, in terms of how it will interrupt the causal mechanisms which perpetuate and sustain a problem in the context being investigated, rather than simply a description of its form.[24] It also requires a deep understanding of the system into which the intervention will be placed to consider 'fit' between intervention and context, how the intervention will interact with context to bring about, aid or sustain the hoped for change, and the likely consequences of attempting to change the context through introducing a new intervention. The realist evaluation approach[25] is one, among several, that strongly emphasise the interaction between context and intervention, contending that programmes have differential effects because the mechanisms by which an intervention influences outcomes may not be activated in all contexts, or may differ from one context to the other.

Understanding the role of contextual influences

The term 'context' has its etymological roots in the Latin *contextus*, meaning 'joining together'. Understanding what happens when a particular intervention is joined together with an individual, team, organisation or health system remains a critical challenge both for science and for practice and policy. But the need to understand and give accounts of context is continually challenged by the problem of conceptualising what 'context' means in practice.[3,6] A taxonomy that could be used across multiple studies to identify and describe critical components of interventions and their contexts might be helpful,[3] as might a synthesis of evidence from a range of disciplines to identify those aspects of context that are most consistently related to intervention success and sustainability.[26] Some inroads have been made to this end in implementation science by Damschroder *et al.*,[27] who synthesised constructs from existing implementation theories to create a framework comprising five key domains (intervention characteristics, outer setting, inner setting, characteristics of the individuals involved and process of implementation) that can be used to guide choices about relevant aspects of the intervention, context and process of implementation to assess in any given evaluation.

To date, many intervention studies lack even basic information about the research and clinical contexts in which they are undertaken,[8] even though item 4b of the CONSORT statement (setting and location)[7] states that researchers should report on 'other aspects of the setting (including the social, economic, and cultural environment and the climate) [that] may also affect a study's external validity'.[7] A specific extension to reporting guidelines to include more detail on context has been proposed.[3,6] Data produced by such reporting may enable the identification of candidate factors suitable for quantitative measurement and the modelling of contextual variables[28] that may be important in revealing variations and inequalities.

This kind of work may also be especially useful in exploring how structural features of contexts may moderate the mechanism through which an intervention produces its effects. For example, a cluster randomised trial of breakfast clubs delivered on a universal basis[29] found that the provision of breakfast at school had no overall impact on rates of breakfast skipping. However, further analysis found that schools in more deprived areas showed an increase in the number of healthy foods eaten at breakfast and found that rates of breakfast skipping were reduced at these schools, suggesting that context moderated the impact of the intervention. Similarly, a peer-led smoking prevention programme aiming to prevent smoking uptake among young people had a greater effect in areas characterised by dense social networks that supported the more efficient diffusion of smoking prevention messages.[30]

The relationship between context and intervention is not, of course, unidirectional. Increasingly, context is seen as having effects so powerful that it may 'shape or co-construct complex interventions and therefore cannot be considered separately from those interventions'.[3] Accordingly, Hawe *et al.*[23] conceive of interventions as events in systems that either 'leave a lasting footprint or wash out depending on how well the dynamic properties of the system are harnessed'. To this end, qualitative research during the earlier stages of intervention development can be used to give insight into the context into which an intervention will be delivered, enabling researchers to better articulate what aspects of the existing context the intervention attempts to change, and why. Where this process of intentional adaption takes place, work by Jansen *et al.*[31] suggests that researchers may attempt to adapt the context to fit the intervention, rather than fashion the intervention to suit the setting.

Interventions may also 'evolve' over time, shaped by a range of contextual influences and pragmatic considerations.[3,32] Interventions involve attempts to change the functioning of 'complex adaptive systems',[23] and the survival of these systems is dependent on ceaseless adaptation in response to changing internal and external environments. Hence adaptations to interventions are often necessary because aspects of the system or its wider context may change after the study has begun, including changes in the economic or political context or local changes.[3,33] Mutations in an intervention may be highly functional, resulting in an intervention that is better suited to the setting in which it is delivered, although adaptations may also have unintended consequences.[3]

Fundamentally, the dynamic interplay between intervention and context inherent to many complex interventions means that it is often difficult, and indeed not always helpful, to separate intervention from context.[23] Viewing interventions as attempts to change aspects of their contexts means that they cannot be understood in isolation from those contexts. Yet adaptations have traditionally been viewed simply as methodological shortcomings, particularly in relation to trials.[3] As a result, researchers may describe (often poorly) the intervention that they intended to deliver rather than that which was actually implemented, thus hampering efforts to replicate or build on research findings. Wells *et al.*[3] urge a cultural shift in which the open and honest reporting of issues such as local adaptations and differences in implementation is perceived as an indicator of study strength and research integrity, rather than a study weakness or lack of researcher ability: what is important is not that these adaptations occurred, but that they are 'known, understood, and reported'. The key to this shift in thinking, Hawe *et al.*[34] argue, is the idea that the integrity of complex interventions is defined functionally, rather than compositionally.

As *Figure 7.1* depicts, capturing the complexity and emergent nature of the interaction between context and intervention[35] is likely to require well-conducted process evaluation at all stages of intervention design, implementation and evaluation.

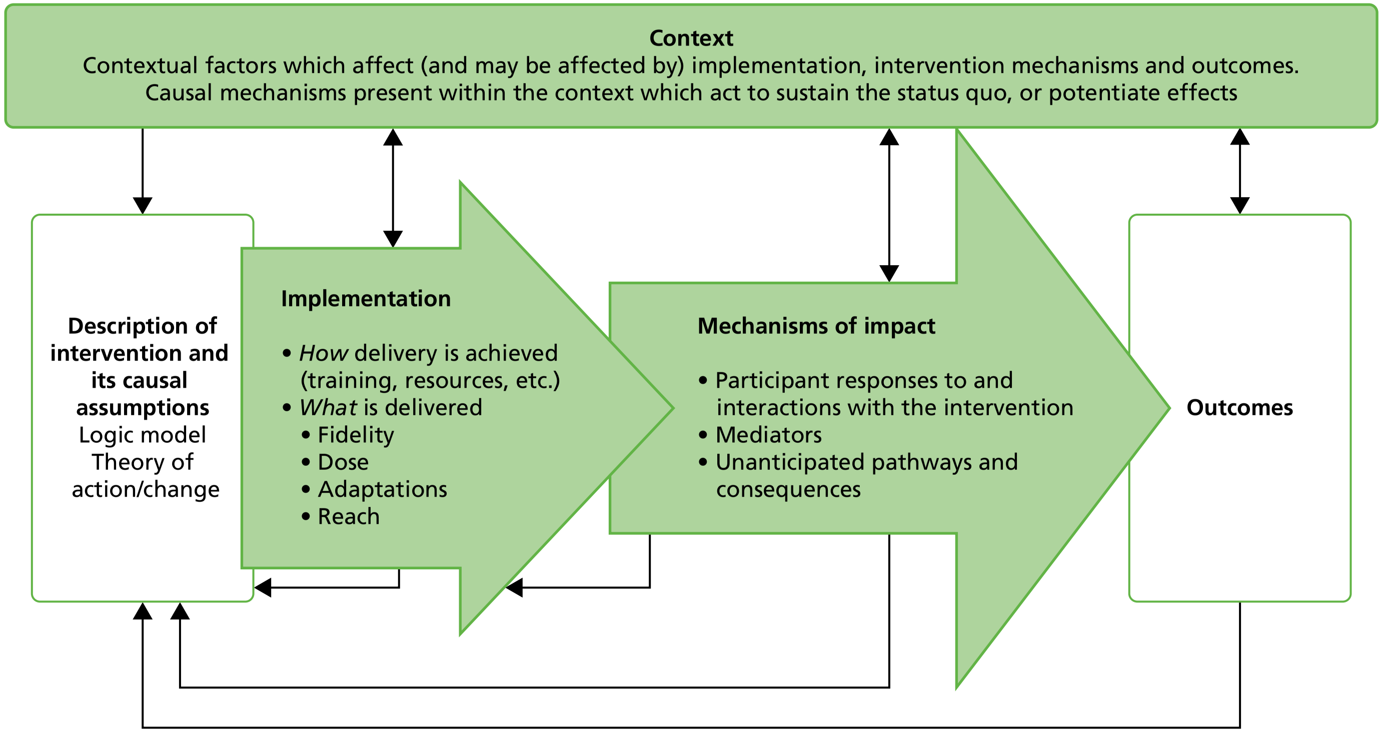
Context
Contextual factors which affect (and may be affected by) implementation, intervention mechanisms and outcomes.
Causal mechanisms present within the context which act to sustain the status quo, or potentiate effects
Description of intervention and its causal assumptions
Logic model
Theory of action/change
Implementation
• How delivery is achieved (training, resources, etc.)
• What is delivered
• Fidelity
• Dose
• Adaptations
• Reach
Mechanisms of impact
• Participant responses to and interactions with the intervention
• Mediators
• Unanticipated pathways and consequences
Outcomes

FIGURE 7.1 Key functions of process evaluation and relationships among them (green boxes represent components of process evaluations, informed by intervention description, which inform interpretation of outcomes). Reprinted from Process Evaluation of Complex Interventions: UK Medical Research Council (MRC) Guidance (p. 24) by Moore *et al*. 2014[35] and used with the permission of the MRC.

The role of qualitative and mixed-methods research

The MRC guidance[1] emphasises the role of qualitative methods during the development and modelling stage that precedes evaluation, thus helping to convince funders of the value of qualitative research.[36] Discussants at the 2015 London meeting felt that the exemplary case studies and synthesis of existing work, like that undertaken by O'Cathain *et al.*[36] and Lewin *et al.*,[37] play an essential role in demonstrating the value that qualitative methods bring. This was seen as especially important for participants, who reported a sense that funders historically have been reluctant to fund development and feasibility work, perhaps especially when policy-makers are eager to get on with implementing a new intervention or policy.

Further mitigating the fruitful use of qualitative research is the emphasis on high-impact publications in assessing the value of institutions, research disciplines and individual researchers. O'Cathain *et al.*[36] found that researchers perceived randomised controlled trials (RCTs) as more likely to facilitate this type of output, with qualitative papers more likely to be published in lower-impact journals. This often resulted in quantitative and qualitative papers being published separately and, in some cases, where qualitative papers were seen as lower-status outputs, not at all.[36] O'Cathain *et al.*[36] suggest a need to address publishing norms that serve as a barrier to the publication of applied mixed-methods research, and for publishers to recognise that in an output-driven environment a lack of high-quality forums for this type of research may have a detrimental influence on the work that is undertaken in the first place. O'Cathain *et al.*[36] also recommend that care is taken to explicitly articulate the role played by the qualitative research and the value brought to the trial (or evaluative study) as well as lessons learned for studies in the future.

An example of this kind of useful work is the study conducted by Redfern *et al.*,[38] who used a range of qualitative methods to develop Stop Stroke, an intervention to improve clinician and patient management of risk factors to prevent stroke recurrence. In establishing the theoretical rationale for the intervention, the research team conducted a systematic review and exploratory interviews to understand patients' experiences of secondary prevention as well as non-participant observation of staff during clinical consultations to understand how prevention advice was given in local contexts; they also mapped patterns of risk factor management practices as well as identifying barriers and facilitators to risk factor management. The resulting programme used these data to design a complex multiple component intervention targeting patients, carers and health-care professionals with tailored advice at three time points post stroke.

This work shows how qualitative methods can be used during the development and modelling phase to design an intervention that takes into consideration the preferences of patients and practitioners, but also takes account of existing local provision. However, this kind of work-up of interventions and contexts remains relatively rare: a number of studies have highlighted the need for more effort to be expended at earlier stages of intervention development and evaluation to maximise the value that qualitative methods in particular can bring.[31,36] These modelling considerations are also discussed in the context of making best use of available evidence in *Essay 1*.

Table 7.1 reports the diverse ways in which qualitative research is used alongside trials to explore various aspects of intervention and study design. The largest subcategory focused on exploring the feasibility and acceptability of the intervention in practice (23%), but relatively little of this work (28%) was undertaken at the pre-trial stage. Only 13% of all papers published between 2008 and 2010 reported pre-trial intervention development work and only one published paper explored the mechanism of action prior to the full trial, suggesting that this important stage of work could be given more emphasis.

Beyond the development phase, qualitative research that runs alongside programme implementation offers a useful way of generating an account of the programme as it happened, rather than as it was planned.[10,28] O'Cathain *et al.*[36] advocate for qualitative research to take a more challenging role, rather than one that is simply descriptive. The programme theory should be specified from the earliest point in the design and development phase[1] but with the expectation that the mechanism that is specified

TABLE 7.1 Framework of the focus of qualitative research used with trials

Category	Subcategory
Intervention content and delivery	Intervention development
	Intervention components
	Models, mechanisms and underlying theory development
	Perceived value and benefits of intervention
	Acceptability of intervention in principle
	Feasibility and acceptability of intervention in practice
	Fidelity, reach and dose of intervention
	Implementation of the intervention in the real world
Trial design, conduct and processes	Recruitment and retention
	Diversity of participants
	Trial participation
	Acceptability of the trial in principle
	Acceptability of the trial in practice
	Ethical conduct of trial
	Adaptation of trial conduct to local context
	Impact of trial on staff, researchers or participants
Outcomes	Breadth of outcomes
	Variation in outcomes
Measures of process and outcome	Accuracy of measures
	Completion of outcome measures
	Development of outcome measures
Target condition	Experience of the disease, behaviour or beliefs

Reproduced from O'Cathain *et al.*[39] under the terms of the Creative Commons Attribution Non-Commercial Licence, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non-commercial and is otherwise in compliance with the licence (http://creativecommons.org/licenses/by-nc/2.0/).[39]

in advance may not necessarily hold true or may need to be adjusted.[10,40] Thus, for example, a study undertaken by Dixon-Woods *et al.*[22] to characterise the Matching Michigan programme in practice involved approximately 855 hours of observational fieldwork in 17 ICUs, including observation of training events, team meetings and external reference group meetings, over 100 interviews with ICU staff and documentary analysis. This enabled them to establish that the programme as implemented was not the same as the one on which it was modelled, and that outcomes varied as a function of outer and inner contextual influences as well as differences in data collection that undermined the comparability of supposedly standardised measures of infection rates across units.[41]

Davidoff *et al.*[40] highlight that aspects of the broader contextual landscape such as cultural, emotional and political challenges, as well as aspects of local context, are all likely to become much more salient as the work proceeds. With this in mind, theoretical accounts of how a programme works should be developed through a continual cycle of theory testing and refinement.[1] To facilitate the enquiring and often iterative nature of qualitative research, O'Cathain *et al.*[36] recommend that ethics boards be more sympathetic to the evolving role of qualitative research throughout a study, allowing researchers to pursue emergent findings without requiring substantial amendments to ethical approvals.

A mapping study carried out by O'Cathain *et al.*[36] that explored the use of qualitative methods alongside RCTs gives some sense of how researchers may employ mixed-methods designs to study complex interventions. The potential value of the qualitative research to the endeavour of generating evidence of effectiveness of health interventions was found to be considerable, and included improving the external validity of trials, facilitating interpretation of the trial findings, improving sensitivity to the human beings who participate in trials, and saving money by steering researchers towards interventions more likely to be effective in future trials.[39] Some researchers perceived qualitative research as peripheral or an 'add-on' to the trial, but others saw the qualitative research as essential and of equal value to the trial. Those researchers who were interested in influencing practice in the real world, and who were aware of the complexity of the intervention they were testing, were particularly likely to see qualitative research as essential; accordingly, they were more likely to attribute more resources to the qualitative research and work within integrated teams. Importantly, they were also more likely to value the integration of the qualitative and quantitative findings, thus enhancing the impact of the qualitative findings on the current or future trial.

These challenges need to be addressed given mounting evidence of the benefits of qualitative research in ensuring success in the design and implementation of interventions and in ensuring adequate accounting for context. Case study methodology, particularly when comparative, may be especially powerful in producing in-depth, multifaceted understandings of a complex issue in its real-life context.[28] For example, a prospective longitudinal ethnographic study undertaken by Hoddinott *et al.*[42] alongside a cluster randomised trial of a group intervention to facilitate breastfeeding in Scotland combined qualitative data with descriptive quantitative data on intervention 'dose' to construct case studies, enabling comparison of policy implementation and impact across seven different localities. The techniques used in Hoddinott *et al.*'s study[42] allowed the development of a model of health service attributes necessary to successfully deliver the policy. For instance, in localities where breastfeeding rates had declined, negative aspects of place (rising deprivation, availability of other group interventions, lack of suitable venues), personnel resources and organisational change predominated, preventing teams from creating or exploiting opportunities for the multidisciplinary partnership working that was crucial for successful implementation. The study design also enabled a full analysis of implementation processes prior to knowing the trial outcomes, thus minimising retrospective selection bias and the kinds of rationalisations that can occur when questions about failure (of the technical intervention or the implementation strategy) are asked in retrospect.[42] As this study shows, the attribution of causality in case studies can be supported by iterative pattern-matching processes that develop explanations, deduce implications of those explanations and seek additional information to check these explanations out. See further considerations regarding implementation in *Essay 8* of this volume.

Their evaluation of the Multi-Payer Medical Home Demonstration, RTI International, The Urban Institute and the National Academy of State Health Policy[33] also employ multimethod case studies with the express purpose of building a detailed picture of the state context and programme structure that can be used to interpret variation in implementation and outcomes. Advanced primary care practices, or 'medical homes', utilise a team approach to care, and aim to improve the quality and co-ordination of health-care services through emphasis on prevention, health information technology, care co-ordination and shared decision-making among patients and their providers. The demonstration will evaluate whether or not advanced primary care practice will reduce unjustified utilisation and expenditures, improve the safety, effectiveness, timeliness and efficiency of health care, increase patient decision-making and increase the availability and delivery of care in underserved areas in participating practices, compared against matched controls. Emerging findings already suggest that context has a bearing on the extent to which practices are able to implement improvement practices. For example, smaller practices and practices in rural areas with provider shortages (e.g. specialists, dentists, mental health providers) and a lack of patient transportation both encountered barriers that practices could not always overcome. Preliminary outcome data suggest some variation across states in health-care utilisation and expenditure, and subsequent investigation will examine whether or not differences at the state, practice and programme levels can account for this. This work highlights how qualitative evidence produced as part of a mixed-methods evaluative study has the potential to be of huge interest to policy-makers, especially where it can be used as a formative mechanism of feedback to augment aspects of the context or intervention that are likely to improve implementation in 'real time'.[1,30,43]

This raises an important point, however, about the role of evaluators in feeding back information during the course of a study. Formative feedback loops may be most easily accommodated during the feasibility and pilot stages of evaluation, but researchers may assume a more passive role in studies that aim to establish effectiveness under real-world conditions to avoid interfering with implementation and changing how the intervention is delivered. Ambiguity about the information that can be expected to be fed back, when, to whom and in what level of detail can create tensions between the different stakeholder groups involved in a collaborative evaluative endeavour.[43,44] It is, therefore, important to establish systems for communicating process information to key stakeholders at the outset of the study, to avoid perceptions of undue interference or that vital information was withheld.[43] To this end Brewster *et al.*[44] have developed a concordat or framework that is designed to be completed collaboratively between evaluation partners with the aim of setting expectations and resolving conflict as it arises during evaluation activity. The process of completing a joint mandate for working appears to be a constructive way of building local ownership and shared vision and emphasising the interdependency of all parties.[44]

Despite the evident value that qualitative evidence brings to the development and evaluation of complex interventions, some participants at the 2015 meeting in London felt that a bias towards favouring quantitative research was pervasive. A number of structural issues were identified by O'Cathain *et al.*[36] as driving the privileging of quantitative over qualitative methods, many of which resonated with the views of researchers participating in the meeting in London. They spoke of a historical climate of underfunding qualitative research, which O'Cathain *et al.*[36] identified as a function in part of researchers either underestimating the cost of the research, or deliberately 'squeezing' the qualitative budget in an attempt to keep the costs of a bid down. At the same time, participants also described the tension created by the unrealistic expectations placed on qualitative research to answer all questions not addressed by quantitative methods,[45] especially where qualitative research is undertaken retrospectively to explain intervention failure.

Conclusions

Context has remained understudied in comparison with interventions. An urgent task is that of deepening the conceptual underpinnings of the interactions between contexts and interventions and developing methods that allow for their exploration and evaluation. In understanding interventions as attempts to change their contexts, theorising about context could, for instance, become the starting point for approaches to intervention and evaluation design that are rooted in understanding of how the problem of interest is created and maintained in a particular community organisation or system.[34] This kind of thinking may help to challenge the currently dominant assumption that the 'best' or the 'ideal' comes from the laboratory and gets progressively compromised in real-world applications, and instead force attention onto the need to avoid simplistic and etiolated accounts of what drives change.[10]

It is now clear that a description of the components of an intervention is not enough to ensure replication and scaling: what matters is likely to be the activation of mechanisms, even if precise activities undertaken to activate those mechanisms differs across contexts. Thinking in this way requires far more attention to the upfront specification and piloting of programme theories and to the characterisation of contexts. It also requires the subsequent updating of these in the light of process evaluations that are explicitly attentive to context.

Poorly reported interventions compound the problems of weak accounting for context. Recent efforts to promote improved reporting includes a checklist to prompt authors to describe interventions (including comparator interventions) in sufficient detail to allow their replication.[11] Comprising 12 items, the Template for Intervention Description to Enable Replication (TIDieR) encourages authors to describe features including which interventions (or control conditions) are delivered to different groups (e.g. materials, procedures, provider, mode of delivery, setting, dose), as well to explain legitimate variants of the intervention as well as unplanned adaptations that occur during the course of delivery. However, further work to support explicit descriptions of context is needed.

Such work should identify the benefits of mixed-method approaches to the study of context. Quantitative measures to explore structural features of context are essential, and more effort in helping to specify these measures and support data collection will be an important focus for the future. Qualitative methods are particularly suited to securing the theoretical basis of interventions, exploring the process by which an intervention or programme is delivered, identifying effects and how they are achieved, and characterising how interventions are adapted in response to contextual influences and contexts themselves are modified by interventions. However, when using mixed methods, researchers need to be sensitive to the fact that qualitative and quantitative research methods have grown out of different paradigms. The strengths of each method should not simply be used to bolster the weakness of the other, but instead should be used as complementary methods to study different phenomena related to the same question.[46,47]

Deepening understanding of the contextual factors that influence the delivery and function of complex interventions will require new norms for the conduct of evaluations of complex interventions. Gathering the rich and varied multimethod data needed to conduct a high-quality evaluation involves close collaboration with intervention developers and implementers. The different goals, perspectives, expectations, priorities and interests, professional languages and norms of practice of these groups may lead to tensions that need to be acknowledged and managed from the outset.

For the future, researchers and funders should be encouraged to think in terms of mixed-methods studies where qualitative and quantitative methods are seen as complementary methods, and to expend greater effort earlier in the development and evaluation process to optimise interventions, identify causal mechanisms and characterise contextual influences. Realisation of these ideas will require funder commitments to adequately resource development work and process evaluations that run alongside trials and other types of outcomes-focused evaluations, more realistic expectations about the time frame for developing an intervention, and journals that are willing to publish the results of exploratory work.

Despite current challenges, there was a sense of an improving climate in which qualitative work was valued. O'Cathain *et al.*[36] highlighted the explicit value placed on qualitative methods and mixed-methods studies in the MRC evaluation framework,[1] and the inclusion of qualitative and mixed-methods researchers on research commissioning panels as key drivers of this positive shift. Participants in the London meeting proposed that current calls for examples of impact (by research councils, Research Excellence Framework, the National Institute for Health Research) should be used to further highlight the contribution of qualitative research and the value of mixed-methods studies.

Finally, it is clear that institutional features of the research landscape – including funding structures, publishing models, and the norms, values and practices of the researcher community and the users of evidence – need critical scrutiny in order that they support contextually sensitive research that is capable of determining if and how complex interventions have their desired effects on health and social care services and systems, and, ultimately, on the people who use them.

Acknowledgements

Contributions of authors

Emma Howarth (Senior Research Associate, Public Health) wrote the first draft of the essay and contributed to the production of subsequent drafts.

Kelly Devers (Senior Fellow, Health Policy) commented on the draft essay.

Graham Moore (Senior Lecturer, Social Sciences and Health) edited the essay and provided additional material.

Alicia O'Cathain (Chair in Health Services Research, Applied Health Research) edited the essay and provided additional material.

Mary Dixon-Woods (Professor of Medical Sociology) contributed to the writing of the first draft of the essay and produced the final draft.

References

1. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. *Developing and Evaluating Complex Interventions: New Guidance*. London: MRC; 2008. URL: http://discovery.ucl.ac.uk/103060/ (accessed 10 April 2016).
2. Moore G. *Understanding the Functioning of Complex Interventions in Context: The Role of Process Evaluation*. Paper presented at Evaluation London, 29 June 2015. 2015.
3. Wells M, Williams B, Treweek S, Coyle J, Taylor J. Intervention description is not enough: evidence from an in-depth multiple case study on the untold role and impact of context in randomised controlled trials of seven complex interventions. *Trials* 2012;**13**:95. http://dx.doi.org/10.1186/1745-6215-13-95
4. Goodin R, Tilly C. *The Oxford Handbook of Contextual Political Analysis*. Oxford: Oxford University Press; 2006. http://dx.doi.org/10.1093/oxfordhb/9780199270439.001.0001
5. Falleti TG, Lynch JF. Context and causal mechanisms in political analysis. *Comp Polit Stud* 2009;**42**:1143–66. http://dx.doi.org/10.1177/0010414009331724
6. Datta J, Petticrew M. Challenges to evaluating complex interventions: a content analysis of published papers. *BMC Public Health* 2013;**13**:568. http://dx.doi.org/10.1186/1471-2458-13-568
7. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, *et al.* CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c869.
8. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? *BMJ* 2008;**336**:1472–4. http://dx.doi.org/10.1136/bmj.39590.732037.47
9. Hoffmann TC, Erueti C, Glasziou PP. Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials. *BMJ* 2013;**347**:f3755. http://dx.doi.org/10.1136/bmj.f3755
10. Dixon-Woods M, Bosk CL, Aveling EL, Goeschel CA, Pronovost PJ. Explaining Michigan: developing an ex post theory of a quality improvement program. *Milbank Q* 2011;**89**:167–205. http://dx.doi.org/10.1111/j.1468-0009.2011.00625.x
11. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, *et al.* Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;**348**:g1687.

12. Osborne NH, Nicholas LH, Ryan AM, Thumma JR, Dimick JB. Association of hospital participation in a quality reporting program with surgical outcomes and expenditures for medicare beneficiaries. *JAMA* 2015;**315**:496–504. http://dx.doi.org/10.1001/jama.2015.25

13. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull* 1996;**119**:254–84. http://dx.doi.org/10.1037/0033-2909.119.2.254

14. Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. Med Care 2009;**47**:356–63. http://dx.doi.org/10.1097/MLR.0b013e3181893f6b

15. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, *et al.* Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012;**6**:CD000259.

16. Power M, Tyrrell PJ, Rudd AG, Tully MP, Dalton D, Marshall M, *et al.* Did a quality improvement collaborative make stroke care better? A cluster randomized trial. *Implement Sci* 2014;**9**:40. http://dx.doi.org/10.1186/1748-5908-9-40

17. Mittman BS. Creating the evidence base for quality improvement collaboratives. *Ann Intern Med* 2004;**140**:897–901. http://dx.doi.org/10.7326/0003-4819-140-11-200406010-00011

18. Schouten LM, Grol RP, Hulscher ME. Factors influencing success in quality-improvement collaboratives: development and psychometric testing of an instrument. *Implement Sci* 2010;**5**:84. http://dx.doi.org/10.1186/1748-5908-5-84

19. Carter P, Ozieranski P, McNicol S, Power M, Dixon-Woods M. How collaborative are quality improvement collaboratives: a qualitative study in stroke care. *Implement Sci* 2014;**9**:32. http://dx.doi.org/10.1186/1748-5908-9-32

20. Provonost P, Needham D, Sean B, Sinopoli D, Chu H, Cosgrove S, *et al.* An intervention to decrease catheter-related bloodstream infections in the ICU. *N Engl J Med* 2006;**355**:2373–83.

21. Bosk CL, Dixon-Woods M, Goeschel CA, Pronovost PJ. Reality check for checklists. *Lancet* 2009;**374**:444–5. http://dx.doi.org/10.1016/S0140-6736(09)61440-9

22. Dixon-Woods M, Leslie M, Tarrant C, Bion J. Explaining Matching Michigan: an ethnographic study of a patient safety program. *Implement Sci* 2013;**8**:70. http://dx.doi.org/10.1186/1748-5908-8-70

23. Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. *Am J Community Psychol* 2009;**43**:267–76. http://dx.doi.org/10.1007/s10464-009-9229-9

24. Hawe P. Lessons from complex interventions to improve health. *Ann Rev Public Health* 2015;**36**:307–23. http://dx.doi.org/10.1146/annurev-publhealth-031912-114421

25. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage Publications Ltd; 1997.

26. Robert G, Fulop N. The Role of Context in Successful Improvement. In *Perspectives on Context*. London: The Health Foundation; 2014. pp. 33–58.

27. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009;**4**:50. http://dx.doi.org/10.1186/1748-5908-4-50

28. Dixon-Woods M. The Problem of Context in Quality Improvement. In *Perspectives on Context*. London: The Health Foundation; 2014. pp. 88–101.

29. Moore G, Murphy S, Chaplin K, Lyons RA, Atkinson M, Moore L. Impacts of the Primary School Free Breakfast Initiative on socio-economic inequalities in breakfast consumption among 9–11-year-old schoolchildren in Wales. *Public Health Nutr* 2014;**17**:1280–9. http://dx.doi.org/10.1017/S1368980013003133

30. Campbell R, Starkey F, Holliday J, Audrey S, Bloor M, Parry-Langdon N, *et al.* An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *Lancet* 2008;**371**:1595–602. http://dx.doi.org/10.1016/S0140-6736(08)60692-3

31. Jansen YJFM, Foets MME, de Bont AA. The contribution of qualitative research to the development of tailor-made community-based interventions in primary care: a review. *Eur J Public Health* 2010;**20**:220–6. http://dx.doi.org/10.1093/eurpub/ckp085

32. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review – a new method of systematic review designed for complex policy interventions. *J Health Serv Res Policy* 2005;**10**(Suppl. 1):21–34. http://dx.doi.org/10.1258/1355819054308530

33. RTI International, The Urban Institute, National Academy of State Health Policy. *Evaluation of the Multi-Payer Advanced Primary Care Practice (MAPCP) Demonstration: First Annual Report*. Washington, DC: Patient-Centered Primary Care Collaborative; 2015.

34. Hawe P, Shiell A, Riley T. Complex interventions: how 'out of control' can a randomised controlled trial be? *BMJ* 2004;**328**:1561–3. http://dx.doi.org/10.1136/bmj.328.7455.1561

35. Moore G, Audrey S, Barker M, Bonell C, Hardeman W, Moore L, *et al.* Process *Evaluation of Complex Interventions: UK Medical Research Council (MRC) Guidance*. London: MRC Population Health Science Research Network; 2014.

36. O'Cathain A, Thomas KJ, Drabble SJ, Rudolph A, Goode J, Hewison J. Maximising the value of combining qualitative research and randomised controlled trials in health research: the QUAlitative Research in Trials (QUART) study – a mixed methods study. *Health Technol Assess* 2014;**18**(38). http://dx.doi.org/10.3310/hta18380

37. Lewin S, Glenton C, Oxman AD. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. *BMJ* 2009;**339**:b3496. http://dx.doi.org/10.1136/bmj.b3496

38. Redfern J, Rudd AD, Wolfe CD, McKevitt C. Stop Stroke: development of an innovative intervention to improve risk factor management after stroke. *Patient Educ Couns* 2008;**72**:201–9. http://dx.doi.org/10.1016/j.pec.2008.03.006

39. O'Cathain A, Thomas KJ, Drabble SJ, Rudolph A, Hewison J. What can qualitative research do for randomised controlled trials? A systematic mapping review. *BMJ Open* 2013;**3**:e002889. http://dx.doi.org/10.1136/bmjopen-2013-002889

40. Davidoff F, Dixon-Woods M, Leviton L, Michie S. Demystifying theory and its use in improvement. *BMJ Qual Saf* 2015;**24**:228–38. http://dx.doi.org/10.1136/bmjqs-2014-003627

41. Dixon-Woods M, Leslie M, Bion J, Tarrant C. What counts? An ethnographic study of infection data reported to a patient safety program. *Milbank Q* 2012;**90**:548–91. http://dx.doi.org/10.1111/j.1468-0009.2012.00674.x

42. Hoddinott P, Britten J, Pill R. Why do interventions work in some places and not others: a breastfeeding support group trial. *Soc Sci Med* 2010;**70**:769–78. http://dx.doi.org/10.1016/j.socscimed.2009.10.067

43. Moore G, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, *et al. Process Evaluation of Complex Interventions*. London: MRC Population Health Science Research Network; 2014. pp.19–45; 64–75. URL: http://decipher.uk.net/wp-content/uploads/2014/11/MRC-PHSRN-Process-evaluation-guidance.pdf (accessed 10 April 2016).

44. Brewster L, Aveling E-L, Martin G, Tarrant C, Dixon-Woods M, Barber N, *et al.* What to expect when you're evaluating healthcare improvement: a concordat approach to managing collaboration and uncomfortable realities. *BMJ Qual Saf* 2015;**24**:318–24. http://dx.doi.org/10.1136/bmjqs-2014-003732

45. Munro A, Bloor MJ. Process evaluation: the new miracle ingredient in public health research? *Qual Res* 2010;**10**:699–713. http://dx.doi.org/10.1177/1468794110380522

46. Sale JEM, Lohfeld LH, Brazil K. Revisiting the quantitative-qualitative debate: implications for mixed-methods research. *Qual Quant* 2002;**36**:43–53. http://dx.doi.org/10.1023/A:1014301607592

47. Blackwood B, O'Halloran P, Porter S. Review: on the problems of mixing RCTs with qualitative research: the case of the MRC framework and the evaluation of complex healthcare interventions. *J Res Nurs* 2010;**15**:511–21. http://dx.doi.org/10.1177/1744987110373860

Essay 8 Challenges for implementation science

Chrysanthi Papoutsi,[1,2] Ruth Boaden,[3,4] Robbie Foy,[5]
Jeremy Grimshaw[6] and Jo Rycroft-Malone[7]

[1]Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[2]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Northwest London, Imperial College London, London, UK
[3]Alliance Manchester Business School, University of Manchester, Manchester, UK
[4]National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) Greater Manchester, Manchester, UK
[5]Academic Unit of Primary Care, Leeds Institute of Health Sciences, Faculty of Medicine and Health, University of Leeds, Leeds, UK
[6]Clinical Epidemiology Program, Ottawa Hospital Research Institute and Department of Medicine, University of Ottawa, Ottawa, ON, Canada
[7]School of Healthcare Sciences, Bangor University, Bangor, UK

Declared competing interests of authors: Ruth Boaden is a member of the Researcher-Led Commissioning Board for the National Institute for Health Research (NIHR). Jeremy Grimshaw reports grants from Canadian Institutes of Health Research and held a Canada Research Chair, during the conduct of the work reported here. Jo Rycroft-Malone is the current Director of the NIHR Health Services and Delivery Research programme.

Published May 2016
DOI: 10.3310/hsdr04160-121

This essay should be referenced as follows:

Papoutsi C, Boaden R, Foy R, Grimshaw J, Rycroft-Malone J. Challenges for implementation science. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 121–32.

List of abbreviations

A&F	audit and feedback
ASPIRE	Action to Support Practices Implementing Research Evidence
CLAHRC	Collaboration for Leadership in Applied Health Research and Care

Abstract

As the field of implementation science comprises a range of different ontological and disciplinary orientations, there is not always consensus on the boundaries of the research agenda or the challenges faced. Explicit discussion about the terminology used in the field would be useful to clarify differences in perspective and conceptual foundations. There is a need to generate an in-depth understanding of intervention design and development by considering the role of theory, the influence of context and meaningful involvement of relevant end-users. Theory-driven, pragmatic evaluation designs are proposed as a solution to produce evidence of intervention effects, and the potential of 'implementation laboratories' is also discussed. A balance has to be sought, however, between rigidly controlled studies and adaptive evaluations that allow emergent changes to the intervention or the implementation process through timely feedback. Practical recommendations would support the development of the field in evaluating the implementation of evidence-based practice.

Scientific summary

Implementation science comprises a range of methodologies, frameworks and theories utilising a variety of assumptions and terminology. The challenges inherent within different approaches were debated in the session on 'Interventions to disseminate and implement evidence-based practice' and have been summarised here.

Participants at the London meeting in 2015, described in the *Introduction* to this volume, debated the need for an overarching definition to establish the scope of implementation science as a field of enquiry. One definition suggested implementation science as encompassing the 'study of theories, process, models and methods of implementing evidence-based practice'. There was also discussion on a more strategic function of the term as a 'boundary object' to communicate different meanings across disciplines and practice communities. Differences between complicated and complex interventions were particularly highlighted, acknowledging the role of complexity, non-linearity and unpredictability.

Discussions at the meeting also focused on the importance of clearly defining the problem being addressed, in order to design rigorous evaluations that ask the right questions while maximising implementation pragmatism. The relevance of a 'pragmatic' approach was deemed necessary in terms of both intervention design and implementation.

The role of theory-based intervention design and development was extensively considered. There was consensus on the significant challenges around how theory becomes applied, when and to what end. It was argued that using conceptually transferable mechanisms to explain interventions and how these might work differently in different contexts and across time could improve transferability, adaptation and scale-up.

Participants at the meeting further debated the need for providing timely feedback and 'good enough' evidence on the effect of interventions in practice. Although, on the one hand, the adoption of more iterative, adaptive and incremental designs may enable the incorporation of formative evaluation and feedback, this could also limit the power of studies designed on the basis of statistically measured effects.

The importance of paying attention to context was mentioned throughout the session, along with consideration of the ontological and epistemological assumptions that underlie different notions of how context influences intervention design and implementation. Other topics included the need to tailor user engagement to the problem at hand to ensure relevance, and understanding the role of researchers at the intersection of academia and clinical practice.

Consistent messages were identified at the London meeting about the need for a theory-driven understanding of intervention design, implementation and evaluation. Practical recommendations about how and when to use the range of evaluation options available would further support the development of the field and support health service improvement.

Background

Debates on challenges for the implementation and dissemination of evidence-based practice have been inherently fragmented within different fields and particular areas of expertise. The increasing emphasis on implementation, through both the Medical Research Council framework for complex interventions[1] and the more recent guidance on process evaluation,[2] has been accompanied by a rapid and broad development of a range of approaches, frameworks and theories utilising a variety of assumptions and terminology. Although there are some helpful reviews of the area,[3,4] there are few widely accepted approaches to implementing evidence-based practice and evaluating implementation. This lack of clarity has proven a barrier to considering the bigger picture of intervention design, implementation, development and evaluation in a comprehensive manner. It has also hindered the development of recommendations that recognise the diverse interests of health-care and implementation communities, while being directly relevant to researchers and policy-makers.

This essay presents a summary of the presentations and discussions from the session on implementing and disseminating evidence-based practice that took place at the meeting in London in 2015 described in the *Introduction* to this volume (referred to below as the London meeting). Key themes include delineating the scope of implementation science as a field of enquiry and conceptualising what constitutes complex interventions in this context; designing rigorous evaluations that ask the right questions while maximising pragmatism of implementation in the real world of practice; using theory-based intervention design and development methodologies; providing timely and appropriate feedback; adequately studying context and engaging in user involvement; and understanding the challenges of scaling up implementation interventions and their evaluation. Participants also reflected on the role of researchers in implementation.

Terminology and definitions

What is implementation science?

A key area of contention at the London meeting concerned the need (or otherwise) for an overarching definition of implementation science to clarify the subject matter of the field. One of the participants argued that a strict or constraining definition might not serve a useful purpose. Instead, it was suggested that a certain level of flexibility in the way 'implementation science' is understood could contribute to the term acting as a 'boundary object'[5] between different communities of research and practice. This flexibility would allow an organising vision to be developed which could potentially pull people together in the direction of a shared understanding.

Attempts were made to combine different views, with the proposal that implementation science be defined as the 'study of theories, process, models and methods of implementing evidence-based practice'. Apart from considering the effectiveness of interventions in terms of clinical outcomes, the aim of implementation science would be to research approaches and methods for closing the gap between what 'we think we know' from evidence, and what is routinely practised. It was acknowledged that the subject matter of implementation science is multifaceted, ranging from the introduction, integration, adaptation and sustainability of innovations, to user engagement, sense-making and division of labour. Intended and unintended consequences of adoption and adaptation would need to be considered in a recursive relationship with intervention design (telecare was mentioned as an illustrative example).

The discussion on terminology at the London meeting extended to consider differences between terms such as implementation and improvement science. It was argued, however, that an extensive discussion on terminology might be esoterically serving the interests of scientists/academics, rather than solving key underlying issues and being of direct relevance to those trying to evaluate how best to close the gap between evidence and practice and to those implementing improvement projects 'on the ground'. Countering that, it might be important to clarify terminology in order to generate a better understanding of epistemological and ontological underpinnings. It was also deemed necessary to develop a common language to engage clinicians. The word 'science' was considered useful in establishing credibility and signifying the authority required for the field to be recognised in the medical world. However, it can also be argued that the term implementation 'research' might better signify the substance and mission of this area, as it does not imply that a new discipline is created. The group concluded that it would be useful if people at least explained the terms they are using, especially if coming from different disciplines and ontological positions.

One of the presentations at the meeting further elaborated on the subject matter of implementation science, by reflecting on different conceptualisations of the 'gap' between evidence and practice. Drawing on Best and Holmes,[6] three approaches were discussed: linear (i.e. knowledge transfer), relationship (i.e. partnerships and coproduction) and system models (i.e. implementation being one component of how an organisation develops/changes). By conceptualising the gap in different ways, the evaluations of interventions are built on different underlying epistemologies and assumptions, and differing relationships with stakeholders. This underpinning frame of reference should be taken in account when evaluating interventions to implement and disseminate evidence-based practice.

What do we mean by complex interventions?

The term 'complex interventions' was another area of contention at the meeting, primarily around conceptual clarity and consistency of use. There was a range of views about the term, with some favouring more 'interrogation' of its meaning, while others provided a more definite picture of what complex interventions comprise. Reference was made to a presentation on randomised controlled trials earlier in the day that defined interventions (1) as complicated to place emphasis on the internal variability of the task (e.g. building a car); and (2) as complex to encapsulate their relationship and variable interactions with context (e.g. raising a child is complex because of the uncertainty and unpredictability of social behaviour). These issues are picked up in a number of other essays emerging from the meeting and appearing in this volume (see *Essays 2*, *3*, *6* and *7*).

The role of complexity, non-linearity and unpredictability was particularly highlighted and many acknowledged the need for a more in-depth understanding of change mechanisms, or underlying theories of change, when interventions have fuzzy boundaries and their purpose is not clearly defined. This has implications for the methodologies used to study implementation. One of the participants at the meeting also noted how complexity emerges when spreading change that does not directly fit with existing culture. Different levels of complexity were illustrated by discussing the example of shared care between patients and health professionals as an intervention where multiple feedback loops and power boundaries might lead to variable outcomes. By contrast, clinical interventions, such as changing the boundaries for reporting abnormality in liver function tests, were deemed to be less complex as their rationale would be less challenging for local health-care cultures.

Intervention design and development

Many of the topics discussed referred to the need for a clear definition of the problem being addressed, the role of theory and the relevance of a 'pragmatic' approach in terms of both intervention design and implementation.

Asking the right question and defining the 'problem'

One of the presentations suggested that intervention design has been one of the weakest elements in implementation science over the last 30 years. Intervention design often happens without consulting previous literature and theory, in what has been characterised as 'an expensive version of trial-and-error'.[7,8]

Before deciding on the optimal design, participants agreed that it was important to fully define the implementation challenge with enough specificity, along with understanding the determinants and barriers to implementing interventions and policies in practice. Specifying the right question should be an interdisciplinary process carried out in conjunction with those implementing the intervention. The right design and appropriate methods would follow from the specification of the question. Discussions also acknowledged a need for finding a middle-ground between resource-intensive interventions that are difficult to implement in practice and 'light' interventions that have little potential for making a difference. This highlights how implementation research and implementation practice are not always viewed as separate, but can be inherently interconnected.

Prospective use of theory was deemed important to identify targets for interventions. However, despite a significant pool of theoretical knowledge available to inform intervention design, how theory becomes applied, when and to what end remains a challenge. The groups at the meeting discussed the importance of establishing what the problem is first (what we need to know) and then looking at which theories might be useful, rather than pre-determining what theory to use. One of the speakers argued for an explicit process of developing interventions, based on an understanding of the determinants of the 'problem' and the perceived mechanism of action of the proposed intervention, as well as logistics and practicalities. Although, pragmatically, implementation researchers are not always involved in prospectively embedding theory in intervention design – for example, where the intervention has been developed by a third party such as a health-care organisation – evaluations can still adopt a theoretically informed approach to explain mechanisms of action.

'Timely and good enough' feedback

Another challenge highlighted was that of providing timely feedback about the effect of interventions and their influence on practice. In cases where intervention studies take years to complete, findings may have diminished relevance to the changing context of the health service. An increased focus on the adoption of more iterative, adaptive and incremental designs may enable the incorporation of formative evaluation and feedback, although the extent of modification of the intervention as a result of this type of feedback gives rise to additional issues about how to make sense of the modifications and their effects. Two examples of projects currently incorporating elements of formative development were given:

- Translating Research in Elder Care (TREC) uses routine data from a minimum data set to provide feedback to care homes piloting interventions and studying the effects on performance (www.kusp.ualberta.ca/Research/ActiveProjects/TREC.aspx).
- A chronic disease management project from the first round of the Greater Manchester Collaboration for Leadership in Applied Health Research and Care (CLAHRC), using Quality and Outcomes Framework information overlaid with a quality improvement intervention, Plan-Do-Study-Act methodology and facilitation, with data fed back on a monthly basis and benchmarked against other sites (http://clahrc-gm.nihr.ac.uk/our-work-2008–2013/chronic-kidney-disease/).

Along with timeliness, the balance between formative and summative evaluation naturally raised questions about what is 'good enough' evidence for service and practice change, what constitutes 'enough' information and at what point would early findings be fed back constructively. It was argued that the potential of generating more implementable evidence at initial trial stage warrants further thought.[9] Discussions seemed to indicate a need for balancing 'rigorous' controlled studies with opportunities to design trials in a way that facilitates the emergence of findings throughout the process. Although stepped-wedge designs were mentioned as a possible methodology, the issues relating to changes in the intervention throughout the process would impact on the statistical power of this type of study, unless it was carefully designed with each step being a separate comparison.

The role of contextual influences was also deemed particularly important, given the need to deliver timely evidence to understand how to embed interventions in the real world. Data accessibility and quality were also considered as issues that might prove challenging in implementation research projects, along with the ability to extract data and using the appropriate analysis tools.

The need to build a cumulative science: the example of audit and feedback

Presentations at the London meeting discussed ways of maximising pragmatism in trials of implementation strategies, referring to audit and feedback (A&F) interventions as an area with a robust evidence base about likely effectiveness of commonly used strategies. The latest Cochrane systematic review of A&F interventions identified 140 randomised trials that, on average, improve care processes (median absolute improvement of +4%), but with a substantial variation in the observed effects (interquartile range of +0.5% to +16%) under specific circumstances or when delivered in certain ways.[10] Metaregression identified some features associated with larger effects from A&F: gaining more information about the context and the nature of the targeted behaviour(s), adding co-interventions or considering the intervention content and delivery strategies.[11] Further analysis identified that no new information on common effect modifiers is emerging from new trials,[12] yet we still need to expand our understanding of how to optimally use A&F.

A pragmatic implementation trial: the ASPIRE example

Policy and practice need to be informed by rigorous evaluations of implementation strategies. As argued by one of the presentations, pragmatic randomised trials generally provide the most valid estimates of 'real-world' intervention effects. To judge the level of pragmatism in trials, the revised PRagmatic Explanatory Continuum Indicator Summary tool (PRECIS-2) proposes nine domains: eligibility criteria, recruitment, setting, organisation, flexibility (delivery), flexibility (adherence), follow-up, primary outcome and primary analysis.[13] As an example meeting these criteria, ASPIRE (Action to Support Practices Implementing Research Evidence) is a National Institute for Health Research-funded programme aiming to develop and evaluate an adaptable implementation package targeting 'high-impact' clinical practice recommendations in primary care (http://medhealth.leeds.ac.uk/info/650/aspire/132/what_is_the_aspire_programme).

Through a process of selecting guideline recommendations, analysing variations in adherence and identifying those with most scope for improvement, the ASPIRE programme involves primary care professionals and patients in developing implementation packages for each of four priorities. The effects of these implementation packages are currently being evaluated in two balanced incomplete block cluster randomised trials. In one trial, 80 practices are randomised to packages to either improve diabetes control or reduce risky prescribing. In the other, 64 practices are randomised to packages to either improve blood pressure control or increase the use of anticoagulation in atrial fibrillation.

However, maximising pragmatism is necessary but insufficient by itself to build a cumulative science of implementation. Theory-guided approaches to both intervention development and evaluation offer advantages in terms of generalisability from using a common conceptual framework, the potential to understand mechanisms of action and the ability to predict consequences in other settings. The role of theory in the study of implementation is further elaborated in *Using theory in implementation science*.

Embedding evaluations within health-care systems: the example of 'implementation laboratories'

'Implementation laboratories' were proposed as a way of utilising existing large-scale service implementation programmes to embed sequential randomised trials that would be testing different ways of delivering implementation interventions in head-to-head comparisons at scale. By being embedded in the health system, implementation laboratories may be useful in clarifying the roles of different stakeholders, creating buy-in and allowing service providers to sustainably develop priorities based on clinical need. This would create a synergistic relationship between clinicians and researchers: the intervention would be developed jointly by the two groups, as would be the interpretation of the results, while clinicians would deliver the intervention and collect data as part of their routine processes. By relying on existing structures, the actual cost of the research would be marginal.

Other benefits for a health system may include it becoming a learning organisation, demonstrable improvements in its quality improvement activities and linkages to academic experts. Implementation science may also benefit from the ability to test important (but potentially subtle) variations in the intervention that may be important effect modifiers. The development of multiple implementation laboratories addressing the same intervention would additionally provide the opportunity to establish a metalaboratory (i.e. a cross-laboratory steering group involving health system participants and internationally leading implementation researchers) to maximise learning (including planned replications and prospective meta-analysis). Examples of existing implementation laboratories include the Affinite programme with the NHS Blood and Transplant programme in the UK[14] and the Ontario Health Implementation Laboratory in Canada.

The role of theory could inform which variations of the intervention should be tested, for example whether or not embedding action/coping plans leads to more effective A&F interventions (based on control theory), or whether or not A&F delivered by credible sources is more effective (drawing on social influence theory). However, variations such as mode of delivery, perceived credibility of source or presenter, frequency of feedback, comparator or level of aggregation would result in a significant number of head-to-head combinations of trials. Follow-up discussions questioned the feasibility of creating intervention laboratories for complicated or complex interventions and noted how theory could be used to identify the necessary elements for intervention design, rather than just guiding which combinations of trials are needed in the first place.[15]

Using theory in implementation science

As is strongly argued in other essays in this volume (particularly *Essays 6* and *7*), theory plays a pivotal role in evaluative methods. Theory-guided approaches to implementation research (rather than purely for intervention design) were considered important for generalisability. It was suggested that using conceptually transferable mechanisms to explain interventions and how these might work differently, in different contexts and across time, can improve transferability, adaptation and scale-up. Theory was deemed to offer a common language to explore and identify influences on practice, enhancing the transparency of intervention development and description.

Which theories?

Phenomena studied by implementation science were deemed multifaceted and dynamic in a way that excludes universal explanations. Combining specific theories to understand different aspects of a problem was suggested as a potentially more productive approach ('horses for courses').

There is a range of different types of knowledge and theoretical perspectives (e.g. middle range, inductive vs. deductive, theory vs. frameworks) with a number of them being used in implementation research and a good summary provided by Nilsen.[4] Examples mentioned included the Theoretical Domains Framework used prospectively in the Translation Research in a Dental Setting (TRiaDS) programme[16] and in ASPIRE (http://medhealth.leeds.ac.uk/info/650/aspire/132/what_is_the_aspire_programme). Some suggested that increased learning could be gained from theories like diffusion of innovations, Normalisation Process Theory,[17] the Promoting Action on Research Implementation in Health Services (PARIHS) framework,[18] the Knowledge to Action framework[19] and the Consolidated Framework for Implementation Research.[20]

Others distinguished between levels of analysis, proposing health and social psychology theories at a more micro level and health services research theories at an organisational level. The usefulness of realist methods and context–mechanism–outcome configurations was also debated. Participants conceded that research should be both problem-driven as well as theory-driven and identified a need to unpack the tensions between the two. They also recognised that theory use can be driven by fads which vary over time. Discussions concluded that the need to use more theory should be accompanied by using it in better ways, not just following the theory path for the sake of it.

The need for developing tools for tracking theoretical fidelity and the integrity of implementation interventions was highlighted. If we accept that interventions change through their enactment with context, then we need to find ways to 'track implementation interventions'. There are a number of existing frameworks which can be used, including the conceptual framework for implementation fidelity[21–24] with ongoing work to expand this described, which includes a learning loop that captures the adaptation processes evolving over time.

Engagement with context and research users

Several features of the context of implementation were discussed at the meeting, along with the role of and potential for engagement with research users.

Studying context

Increasing focus on theory-guided interventions in implementation science led to consideration of the role and influence of context in more depth. One of the presentations outlined different ways of paying attention to context, including natural experiments, such as the CLAHRCs and other relevant cases in the literature.[25] It was also explained how realist methods can elucidate 'what works for whom, how and in what context', as part of implementing interventions which in themselves change as a result of being embedded in particular contexts (capturing generative causation). Examples of realist evaluations have been discussed in the literature,[26] along with debates on how to do realist evaluation within an implementation context, how to identify specific mechanisms and how to demonstrate their interaction with context. The concept of realist trials was also introduced, accompanied by a note on the epistemological and ontological contradictions underlying this approach. This points to relevant debates in the literature that discuss whether randomised controlled trials based on positivist ontological and epistemological assumptions can be synergistically reconciled with a realist approach to context and complexity that draws attention to non-linearity, emergence, adaptation, path-dependence and human agency.[27–29]

One view of context was articulated as consisting of specific elements that need to be taken in account in each setting, such as environment characteristics that differ between, for example, an emergency department and an inpatient ward. By identifying the six most common elements of the context or those that are more likely to have an impact, the ability to scale-up and transfer between contexts can be improved.

User engagement

The importance of user engagement was consistently mentioned, but at the same time it was acknowledged as one of the most challenging aspects of intervention design and implementation. While there was consensus on the importance of user involvement to create appropriately designed interventions, this did not always mean that these interventions would be more 'implementable'. Participants recognised that there is no set recipe for engagement, but this needs to be tailored depending on the problem at hand to ensure relevance. For example, in some cases the end-user might be the patient; in others it might be the hospital Chief Executive Officer who needs to be engaged. Many conceded that the problem should be owned by the clinical community, with the researchers responsible for helping clinicians think through the situation, use appropriate theoretical lenses and contribute with external knowledge. However, the discussion appeared to be based on an unarticulated assumption that 'researchers' are not already clinicians themselves. Difficulties identified included involving the right people (e.g. which constituencies do people represent), especially those who are going to encounter the most barriers during implementation at different stages of the process (design, conduct and evaluation). Participants also recognised the importance of acknowledging the politics of user engagement, in terms of vested interests on the side of practitioners and academics that need to be negotiated and resolved.

Scaling up

The challenge of scaling up interventions was also mentioned, especially regarding how scale-up could be incorporated in evaluation methods and approaches to implementation science in order to improve the chances of successful spread. The role of theory building was again particularly highlighted in supporting implementation across different settings, thus enabling transferability of the intervention. Another aspect of this discussion was the tension between flexibility and standardisation of interventions: how to scale up and at the same time allow adaptation to local structures and increased ownership.

Experiences were described where a lack of understanding about prerequisites to implementation and scale-up had led to interventions having to be deconstructed retrospectively to understand the mechanisms of action. This was especially found to be a limitation when funders of projects decide to spread innovations without a clear understanding of whether or not these are 'implementable' before scale-up.

The role of researchers

Having identified challenges around providing timely and good enough evidence, discussions at the London meeting ensued on the role and responsibilities of researchers as implementation scientists. Participants debated whether or not implementation scientists could act as mediators or brokers responsible for translating research in different settings to ensure relevance between existing studies and the clinical world. It was also suggested that pressure from policy-makers to deliver change in ways that may seem unrealistic could be countered by sharing expertise on what has already been found to be ineffective or should be avoided.

Participants recognised, however, that career incentives in many areas of academia do not always overlap with health service priorities and dissemination in academic journals was deemed to have little impact on policy-making. Some also mentioned how they had encountered difficulties aligning implementation research studies with career paths driven by performance measures of traditional clinical academic disciplines. Overall, there was consensus about the fact that researchers working in the area have to reconcile career incentives and structures that may often seem conflicting (high-impact, high-quality research and direct benefit to patients in practice).

Implications

This summary indicates the need for appropriate communication of the aims and objectives of implementation science to a wider audience. As the field is constituted through the contribution of different ontological and disciplinary perspectives, consensus on the boundaries of the research agenda for implementation scientists might be useful, with some agreement on terminology and conceptual foundations, such as the meaning of complex interventions.

Consistent messages were identified throughout the session at the London meeting about the need for a clear understanding of intervention design and development on the basis of engaging end users and adequately considering contextual influences. Theory-driven, pragmatic evaluation designs were suggested as suitable to producing evidence of intervention effects, along with a consideration of the potential of implementation laboratories. A balance was sought between evaluations that allow emergence through timely feedback and more rigidly controlled studies where the 'intervention' itself does not change (adaptive vs. fixed designs). Persisting limitations with current application of theory remain a challenge towards developing transferable concepts or building what was described as 'cumulative science'.

Practical recommendations that can be meaningfully adopted across research and implementation communities would support the future development of the field in implementing evidence-based practice to improve patient care and outcomes in health-care settings. Implementation scientists operating at the intersection of academic enquiry and practical application may need to be better equipped and supported to respond to what may often be viewed as conflicting priorities.

Acknowledgements

We would like to thank the session co-chairs for facilitating the breakout discussions (Sue Mawson, Julie Reed and Paul Wilson) and the note-takers for providing us with material for this essay (Elizabeth Gibbons and Simon Turner).

Contributions of authors

Chrysanthi Papoutsi (Health Services Researcher) summarised the presentations and discussions held during the session on 'Interventions to disseminate and implement evidence-based practice', drafted and revised this essay.

Ruth Boaden (Professor, Implementation Science) contributed to the writing of the essay, provided comments and suggested additional references.

Robbie Foy (Professor, Primary Care), **Jeremy Grimshaw** (Professor, Implementation Science and Health Services Research) and **Jo Rycroft-Malone** (Professor, Implementation Science and Health Services Research) reviewed the essay and provided comments.

References

1. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;**337**:a1655.
2. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, *et al.* Process evaluation of complex interventions: Medical Research Council guidance. *BMJ* 2015;**350**:h1258. http://dx.doi.org/10.1136/bmj.h1258
3. van Achterberg T. Introduction to Section 4: Implementation of Complex Interventions. In Richards DA, Hallberg IR, editors. *Complex Interventions in Health: An Overview of Research Methods*. London: Routledge; 2015. pp. 261–4.
4. Nilsen P. Making sense of implementation theories, models and frameworks. *Implement Sci* 2015;**10**:53. http://dx.doi.org/10.1186/s13012-015-0242-0
5. Star SL, Griesemer JR. Institutional ecology, translations' and boundary objects: amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Soc Stud Sci* 1989;**19**:387–420. http://dx.doi.org/10.1177/030631289019003001
6. Best A, Holmes B. Systems thinking, knowledge and action: towards better models and methods. *Evid Policy* 2010;**6**:145–59. http://dx.doi.org/10.1332/174426410X502284
7. French SD, Green SE, O'Connor DA, McKenzie JE, Francis JJ, Michie S, *et al.* Developing theory-informed behaviour change interventions to implement evidence into practice: a systematic approach using the Theoretical Domains Framework. *Implement Sci* 2012;**7**:38. http://dx.doi.org/10.1186/1748-5908-7-38
8. Eccles M, Grimshaw J, Walker A, Johnston M, Pitts N. Changing the behavior of healthcare professionals: the use of theory in promoting the uptake of research findings. *J Clin Epidemiol* 2005;**58**:107–12. http://dx.doi.org/10.1016/j.jclinepi.2004.09.002
9. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, *et al.* Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. *BMC Med* 2010;**8**:63. http://dx.doi.org/10.1186/1741-7015-8-63
10. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, *et al.* Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev* 2012;**6**:CD000259. http://dx.doi.org/10.1002/14651858.cd000259.pub3
11. Ivers NM, Sales A, Colquhoun H, Michie S, Foy R, Francis JJ, *et al.* No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention. *Implement Sci* 2014;**9**:14. http://dx.doi.org/10.1186/1748-5908-9-14
12. Ivers NM, Grimshaw JM, Jamtvedt G, Flottorp S, O'Brien MA, French SD, *et al.* Growing literature, stagnant science? Systematic review, meta-regression and cumulative analysis of audit and feedback interventions in health care. *J Gen Int Med* 2014;**29**:1534–41. http://dx.doi.org/10.1007/s11606-014-2913-y
13. Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;**350**:h2147. http://dx.doi.org/10.1136/bmj.h2147
14. Gould NJ, Lorencatto F, Stanworth SJ, Michie S, Prior ME, Glidewell L, *et al.* Application of theory to enhance audit and feedback interventions to increase the uptake of evidence-based transfusion practice: an intervention development protocol. *Implement Sci* 2014;**9**:92. http://dx.doi.org/10.1186/s13012-014-0092-1
15. ICEBeRG (Improved Clinical Effectiveness through Behavioural Research Group). Designing theoretically-informed implementation interventions. *Implement Sci* 2006;**1**:4. http://dx.doi.org/10.1186/1748-5908-1-4

16. Clarkson JE, Ramsay CR, Eccles MP, Eldridge S, Grimshaw JM, Johnston M, *et al.* The translation research in a dental setting (TRiaDS) programme protocol. *Implement Sci* 2010;**5**:57. http://dx.doi.org/10.1186/1748-5908-5-57

17. May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, *et al.* Development of a theory of implementation and integration: Normalization Process Theory. *Implement Sci* 2009;**4**:1–9. http://dx.doi.org/10.1186/1748-5908-4-29

18. Rycroft-Malone J. The PARIHS Framework – a framework for guiding the implementation of evidence-based practice. *J Nurs Care Qual* 2004;**19**:297–304. http://dx.doi.org/10.1097/00001786-200410000-00002

19. Wilson KM, Brady TJ, Lesesne C, on behalf of the NCCDPHP Workgroup on Translation. Peer reviewed: an organizing framework for translation in public health: the knowledge to action framework. *Prev Chronic Dis* 2011;**8**:A46.

20. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009;**4**:50. http://dx.doi.org/10.1186/1748-5908-4-50

21. Masterson-Algar P, Burton CR, Rycroft-Malone J, Sackley CM, Walker MF. Towards a programme theory for fidelity in the evaluation of complex interventions. *J Eval Clin Pract* 2014;**20**:445–52. http://dx.doi.org/10.1111/jep.12174

22. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. *Implement Sci* 2007;**2**:40. http://dx.doi.org/10.1186/1748-5908-2-40

23. Hasson H. Systematic evaluation of implementation fidelity of complex interventions in health and social care. *Implement Sci* 2010;**5**:67–75. http://dx.doi.org/10.1186/1748-5908-5-67

24. Hasson H, Blomberg S, Dunér A. Fidelity and moderating factors in complex interventions: a case study of a continuum of care program for frail elderly people in health and social care. *Implement Sci* 2012;**7**:1–11. http://dx.doi.org/10.1186/1748-5908-7-23

25. Scarbrough H, D'Andreta D, Evans S, Marabelli M, Newell S, Powell J, *et al.* Networked innovation in the health sector: comparative qualitative study of the role of Collaborations for Leadership in Applied Health Research and Care in translating research into practice. *Health Serv Del Res* 2014;**2**(13). http://dx.doi.org/10.3310/hsdr02130

26. Salter KL, Kothari A. Using realist evaluation to open the black box of knowledge translation: a state-of-the-art review. *Implement Sci* 2014;**9**:115. http://dx.doi.org/10.1186/s13012-014-0115-y

27. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc Sci Med* 2012;**75**:2299–306. http://dx.doi.org/10.1016/j.socscimed.2012.08.032

28. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Methods don't make assumptions, researchers do: a response to Marchal *et al.* *Soc Sci Med* 2013;**94**:81–2. http://dx.doi.org/10.1016/j.socscimed.2013.06.026

29. Marchal B, Westhorp G, Wong G, Van Belle S, Greenhalgh T, Kegels G, *et al.* Realist RCTs of complex interventions – an oxymoron. *Soc Sci Med* 2013;**94**:124–8. http://dx.doi.org/10.1016/j.socscimed.2013.06.025

Epilogue

Ray Fitzpatrick[1] and Rosalind Raine[2]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, UK
[2]Department of Applied Health Research, University College London, London, UK

Declared competing interests of authors: none

Published May 2016
DOI: 10.3310/hsdr04160-133

This epilogue should be referenced as follows:

Fitzpatrick R, Raine R. Epilogue. In Raine R, Fitzpatrick R, Barratt H, Bevan G, Black N, Boaden R, *et al.* Challenges, solutions and future directions in the evaluation of service innovations in health care and public health. *Health Serv Deliv Res* 2016;**4**(16). pp. 133–36.

Generalisable insights

This collection of essays offers insights from world-leading researchers on important methodological issues in health care and public health evaluation. The authors have stimulated debate as to how these issues might be addressed and proposed thoughtful approaches to tackle often competing priorities such as rigour and the need for prompt results. In doing so, the case for plurality in methodological approaches and for sustained investment in new and diverse methods is well made. Other commonly occurring themes can be drawn out from these essays.

For example, authors provide wise counsel about the need to engage early and often with policy-makers, practitioners, funders and interdisciplinary researchers about potential research. Discussions should be transparent and wide ranging to ensure that common understandings are achieved about perceived objectives, the identification of relevant and feasible outcomes and the maintenance of appropriate boundaries between interested parties. Such discussions will often shift the focus from the traditional binary question of effectiveness towards a broader understanding of mechanisms, processes and outcomes. They might, for example, lead to a greater focus on understanding how to manage major service change, or the interplay between system change, context and time. They can lead to rich and valuable discussions about outcomes of relevance to stakeholders with differing perspectives, priorities and timelines.

The role of theory and of models to predict and explain mechanisms of change was analysed. Any tendency to regard specific theories as inviolable was challenged. However, the value of critically applying and developing theories and models, for the evaluation both of intervention quality and of implementation, is endorsed.

There is also consensus for the use of mixed and multiple methods. The importance of integrating observational research with experimental methods is now recognised. Methods to combine quantitative and qualitative research designs in evidence synthesis are rapidly improving. However, authors went further and provided exemplars of uniting established methods with innovative techniques to enhance the quality and utility of research, such as placebo tests alongside randomised trials to tackle external validity.

These essays include a clear call to co-ordinate analyses of macro-, meso- and micro-level determinants of system change. Research conclusions are incomplete if we do not take account of the interplay between political, structural, economic and organisational factors, with the sociopsychological determinants of behaviour in groups and individuals. Referred to as ‘context’, macro- and meso-level factors are too often dismissed as unique features, too intangible to define and therefore immovable barriers to the successful transplantation of service innovation to other settings. This is unhelpful; instead, the key is to identify generalisable elements of beneficial processes and outcomes by exploring the mutual influence of the intervention with the various, distinctly defined contextual components.

Such extensive analyses require access to health and care data from multiple sources and sectors. This raises issues concerning the urgent need to improve the quality and comprehensiveness of routine data, particularly from non-acute settings and with regard to clinical information such as comorbidity and severity. Although data security is paramount, the lack of progress in giving researchers prompt and appropriate access to these data is placing a genuine strain on the ability to conduct research. The problem is well known and solutions have been exhaustively worked through. Decisive leadership from key policy-makers is now essential to expedite the course of action required.

Finally, we are urged to be ambitious and to innovate by drawing from other disciplines. Applied health research is beginning to benefit from collaborations with computer scientists, spatial analysts and mathematicians. Examples cited include the application of machine learning to code narrative data, computer adaptive testing to reduce response bias and interactive multimedia techniques to examine clinical decision-making. However, it is clear that our colleagues in other fields are making rapid progress in areas of immense value to our research and that our response so far has been spasmodic. We must investigate and embrace opportunities to make methodological leaps which could transform our ability to deliver comprehensive, rigorous and timely evaluations of service change.

What we missed and a way forward

We know that we have missed important topics. We have not, for example, explored the application of artificial intelligence to modelling and analysis, and have barely mentioned relational methods (such as social and organisational network analysis) or qualitative methods such as participatory research and action ethnography. More needs to be said about how best to develop the vital contributions from patients, the public and policy-makers.

No doubt readers will identify other opportunities overlooked which will broaden the debate. However, we never intended to present an exhaustive account of the field. Instead, we hope that these essays reinvigorate dynamic engagement in a fast-moving area by providing authoritative guidance on the best ways forward. They have put paid to any notion that it is necessary to forego excellence in order to deliver prompt results. The opportunity costs of 'quick and dirty' research can be profound. This collection has, therefore, maintained a keen eye on the need to achieve accurate, comprehensive, relevant, timely and generalisable results. We hope that this collection will stimulate further responses and to maintain momentum in developing the field for the benefit of patients and the public.

EME

HS&DR

HTA

PGfAR

PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

This report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health

Published by the NIHR Journals Library