

Practical issues that impact
statistical design, analysis
and synthesis of cluster
randomised controlled trials

A thesis submitted to the University of
Manchester for the degree of Doctor of
Philosophy (PhD by Published Work) in
the Faculty of Biology, Medicine and
Health

Sarah Rhodes

2022

Contents

Abstract.....	4
Declaration.....	5
Copyright.....	7
Statement of eligibility.....	8
Part (i) Personal statement.....	8
Part (ii) A complete and numbered list of the publications submitted.....	11
Acknowledgements.....	13
Chapter 1: Introduction.....	14
1.1 Background.....	14
1.2 Thesis structure.....	14
1.3 Overview of Chapter 2; Design and analysis of CRTs.....	14
1.4 Overview of Chapter 3; Novel methods to incorporate cluster randomised crossover trials in meta-analysis.....	15
1.5 Overview of Chapter 4; Novel methods of meta-analysis to synthesize data from trials of complex interventions with mixed levels of clustering.....	15
2 Chapter 2 Design and analysis of CRTs.....	17
2.1 Chapter overview.....	17
2.2 Pragmatic approach to sample size estimation.....	17
2.2.1 Overview of methods to calculate sample size in CRTs.....	17
2.2.2 Example of a pragmatic approach to sample size calculation; OSCARSS.....	20
2.3 Methods to minimise selection bias.....	28
2.3.1 Evidence of selection bias in CRTs.....	28
2.3.2 Using modified informed consent to minimise selection bias.....	29
2.3.3 Using recruitment before randomisation to minimise selection bias.....	33
2.3.4 Other methods to minimise selection bias.....	33
2.3.5 Methods used in OSCARSS to minimise selection bias.....	33
2.4 Analysis and interpretation.....	36
2.4.1 Overview of methods of analysis for CRTs.....	36
2.4.2 OSCARSS analysis and interpretation.....	38
2.5 Reflections.....	41
2.6 Metrics.....	43
3 Chapter 3 Novel methods to incorporate cluster randomised crossover trials in meta-analysis.....	44

3.1	Chapter overview	44
3.2	Including cluster randomised designs in meta-analysis.....	44
3.2.1	Systematic reviews and meta-analysis	44
3.2.2	Meta-analysis of CRTs	46
3.2.3	Issues in the meta-analysis of CRTs	47
3.3	Alternative cluster designs.....	47
3.3.1	Different designs	47
3.3.2	The CRXO design	48
3.3.3	Including CRXO designs in meta-analysis.....	49
3.4	Methods applied to Cochrane Review of chlorhexidine bathing	49
3.4.1	Background	49
3.4.2	Appropriate estimation of hospital acquired infection rates	50
3.4.3	Design effects in CRXO trials	52
3.4.4	Combining estimates of hospital infection rates	56
3.4.5	Sensitivity analyses	57
3.5	Recommendations for combining CRXO trials in meta-analysis.....	58
3.6	Reflections.....	59
3.7	Metrics	60
4	Chapter 4 Incorporating CRTs into a meta-analysis of complex behaviour change interventions.....	61
4.1	Chapter overview	61
4.2	Systematic Reviews of complex interventions	61
4.2.1	Complex behaviour change interventions	61
4.2.2	Challenges for systematic reviews of complex interventions.....	61
4.2.3	Methods suitable for systematic reviews of complex behaviour change interventions.....	62
4.3	Challenges faced in the SOCIAL review	63
4.3.1	Description of the SOCIAL systematic review	63
4.4	Methods chosen for the SOCIAL systematic review	65
4.4.1	Meta-analysis methods in the SOCIAL systematic review	65
4.4.2	Comparisons in the SOCIAL review	66
4.4.3	Levels of clustering.....	69
4.4.4	Standardised mean differences	71
4.5	Reflections.....	71
4.6	Metrics	72
5	References	74

6	My submitted papers.....	83
6.1	Paper1	84
6.2	Paper 2	85
6.3	Paper 3	86
6.4	Paper 4	87
6.5	Paper 5	88
6.6	Paper 6	89
6.7	Paper 7	90

ABSTRACT

Practical issues that impact statistical design, analysis and synthesis of cluster randomised controlled trials

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy (PhD) in the School on Health Sciences, Faculty of Biology, Medicine and Health

Sarah Rhodes, 2022

When conducting randomised controlled trials assessing the effect of introducing a new health service or policy, researchers face challenges relating to organisation, implementation and contamination. Cluster randomised trials, where participants are randomised in groups, offer a solution to these challenges.

In this PhD thesis I present seven published research articles as evidence of my contribution as an applied statistician to the design, analysis and synthesis of cluster randomised trials.

I present two papers from the Organising Support for Carers of Stroke Survivors (OSCARSS) trial to demonstrate the methods I developed to minimise bias in the design and analysis of a cluster randomised trial of a complex intervention. I present a Cochrane systematic review to display innovative methods that I developed to incorporate trials with non-standard cluster designs, such as cluster crossover trials, into systematic reviews. SOCIAL was a large systematic review of social norms interventions to change healthcare professional behaviour. I present 4 papers relating to this review to illustrate how I overcame a number of challenges which included synthesis of multiple complex interventions, a mixture of outcome measurements, and a variety of cluster designs.

I have shown that cluster randomised trials reveal a number of challenges in their design, analysis and evidence synthesis, over and above those of individually randomised trials. I have offered practical methods to deal with these challenges and critically appraised their merits and weaknesses.

DECLARATION

Candidate Sarah Rhodes

Faculty Medicine

Thesis Title Practical issues that impact statistical design, analysis and synthesis of cluster randomised trials

i The nature and extent of the candidate's own contribution and the contribution of co-authors and other collaborators to each of the publications presented is as follows:

Paper 1

I was co-investigator and lead statistician on this large, national, cluster randomised trial funded by NIHR CLAHRC Greater Manchester. I was involved from the start and made a substantial contribution to the conception of the cluster randomised design, choice of outcome measures and time points, exploration of sample size and practical issues such as methods for consent and data collection. I led the writing of the methods and analysis sections of the protocol and I was lead author on the published Statistical Analysis Plan.

Paper 2

I was co-investigator and lead statistician on this large, multicentre cluster randomised trial. I was heavily involved in design, wrote the statistical analysis plan, was a member of the trial management group, I oversaw data collection and management, I supervised the analysis, designed all statistical tables, and wrote much of methods, analysis and results sections. With Patchwood and Bowen I provided edits and responses in light of reviewer comments.

Paper 3

I took statistical lead for this methodologically challenging review: I developed methods that utilised all of the available information, extracted data, performed the analysis, reported the methods and changes to the protocol in a transparent replicable way, assisted with the GRADE summary and reporting of results.

Paper 4

The initial idea came from Cotterill and then discussions between Cotterill, Powell and myself led to the concept of the review in terms of research questions, inclusion criteria and methods. I was a co-investigator on the NIHR grant, led on the statistical methods, extracted data for a scoping review and was lead author of the methods sections in this protocol.

Paper 5

I was co-applicant on the NIHR funding application for this project. I was lead on the statistical methods. With Cotterill and Tang I designed the methods for study screening, data extraction and risk of bias assessment. I performed a large proportion of the screening, data extraction and risk of bias assessment. I planned and conducted all meta-analyses and network meta-analyses on the primary outcome. I was the lead author of the statistical methods and results sections and was heavily involved in interpretation of results and edits of all other sections.

Paper 6

As above. This journal article summarises the NIHR report above. I worked with Tang and Cotterill to write this concise paper based on the NIHR report. I was heavily involved in edits and responses in relation to reviewer comments.

Paper 7

I was lead author on this methods paper. I had the original idea, wrote the first draft, performed all analyses and coordinated all editing and submission.

ii. All of the work presented has been completed whilst the candidate has been a member of staff of this University;

iii. None of the work presented has been submitted in support of a successful or pending application for any other degree or qualification of this or any other University or of any professional or learned body.

I confirm that this is a true statement and that, subject to any comments above, the submission is my own original work.

Signed:

Date:

COPYRIGHT

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and they have given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see <http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420>) , in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see <http://www.library.manchester.ac.uk/about/regulations/>) and in the University's policy on Presentation of Theses.

STATEMENT OF ELIGIBILITY

Part (i) Personal statement

After an initial career as a Mathematics teacher I studied for an MSc in Statistics (with Distinction) from the University of Manchester in 2003. From 2003-2007, I was employed full time by the Pennine Acute Hospitals NHS Trust as a Medical Statistician. From 2007-2015 I took a career break to care for my young family and to live overseas due to my husband's job.

Upon return to the UK, I began my University research career as a Research Associate with the Centre for Biostatistics in 2015. Since then my career has progressed and my research portfolio and level of responsibility has grown. I was promoted to Research Fellow in 2019, given line management responsibility for a junior colleague.

I see my career very much in multidisciplinary collaborative research, taking the academic lead in statistical issues of research design, analysis and interpretation. As is common for an applied statistician, my involvement in projects has been as Principal Statistician rather than Principal Investigator or first author.

There is an art to meeting new researchers with an idea and trying to tease out of them what their research question really is, and how best to apply the statistical trade-off between precision and bias within the given context.

I have successfully built multi-disciplinary collaboration with several research teams and I have been the co-investigator and lead statistician on a number of substantial research projects, including two NIHR funded trials, two NIHR funded systematic reviews and a stepped wedge trial funded by Greater Manchester Cancer. My intellectual input into these projects has been critical in securing external funding. I have helped to design studies that best answer the research questions while minimising bias, and I have helped to ensure timely delivery of these projects. As well as leading on data management and analysis, I have contributed to management groups and been heavily involved in paper writing.

I have promoted a model of statistical collaboration that ensures involvement throughout, and fully utilises the statistician's knowledge of trial design and outcome measurement; this is not always easy as many researchers still feel that the role of the statistician is limited to sample size and analysis.

In response to the COVID-19 pandemic, I have spent the last 18 months working on the UK Health and Safety Executive funded PROTECT programme (Partnership for Research in Occupational, Transport, Environmental COVID Transmission <https://sites.manchester.ac.uk/covid19-national-project/research-themes/sector-specific-studies/>) where I co-lead one of the work packages; I lead analyses of epidemiological datasets and collaborate with other colleagues with their research plans and analyses. From this work I have six published peer-reviewed papers (including one as first author (1), and more papers in draft. In Jan 2022 I obtained my first grant as Principal Investigator from the Office for National Statistics, leading a team of statisticians working on questions relating to occupation and COVID-19 using the ONS Coronavirus (COVID-19) Infection Survey.

Degrees

September 2004	M.Sc. (with Distinction) in Statistics, University of Manchester, UK
July 1996	Post Graduate Certificate of Education, University of Oxford, UK
July 1995	B.Sc. in Applied Mathematics (II.I), University of Warwick, UK

Employment History

Aug 2019 – present	Research Fellow, Centre for Biostatistics, University of Manchester.
Sept 2015 – Aug 2019	Research Associate, Centre for Biostatistics, University of Manchester.
April 2012–Sept 2013	Part-time Statistics Advisor for The School of Tourism, University of Queensland, Brisbane, Australia
Oct 2008 – June 2009	Part-time Mathematics Teacher, New Cairo British International School, Egypt
Sept 2004 – Aug 2007	Medical Statistician for Pennine Acute Hospitals National Health Service (NHS) Trust and Salford Royal Hospitals NHS Trust, UK
Jan 2003 - July 2003	Mathematics Teacher for Cheshire County Council, UK (Maternity cover)
Oct 2000 – Nov 2002	Advisor for Mathematics for the University of Namibia, with Voluntary Services Overseas
Sept 1996 - July 2000	Mathematics Teacher for Wolverhampton Borough Council, UK

Research

Frontline adviser for NIHR Research Design Service

Author of 24 peer-reviewed journal articles

Secured £1.7 million of external research funding as a co-investigator

Principal Investigator on Office of National Statistics funded grant (£64,732)

Division Ethics Signatory

Senior Statistical Editor for Cochrane Gut Group

Statistical Editor, Journal of Maternal and Child Nutrition

Teaching

Tutor and marker on the Evidence Based Practice module of the Masters in Population Health.

Statistical support to students at undergraduate, Masters and PhD level.

Lecturer on 'Introduction to RCTs' on the African Institute of Mathematics (AIMS) programme in Cameroon.

Statistical lead on Evidence Based Medicine course for undergraduate medical students.

Part (ii) A complete and numbered list of the publications submitted

Paper 1

Patchwood E, Rothwell K, Rhodes S, Batistatou E, Woodward-Nutt K, Lau Y-S, Grande G, Ewing G, Bowen A. Organising Support for Carers of Stroke Survivors (OSCARSS): study protocol for a cluster randomised controlled trial, including health economic analysis. *Trials*. 2019;20(1):19. <https://doi.org/10.1186/s13063-018-3104-7>

Paper 2

Patchwood E, Woodward-Nutt K, Rhodes SA, Batistatou E, Camacho E, Knowles S, Darley S, Grande G, Ewing G, Bowen A. Organising Support for Carers of Stroke Survivors (OSCARSS): a cluster randomised controlled trial with economic evaluation. *BMJ Open*. 2021;11(1):e038777. <http://dx.doi.org/10.1136/bmjopen-2020-038777>

Paper 3

Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. *Cochrane Database of Systematic Reviews*. 2019(8). <https://doi.org/10.1002/14651858.CD012248.pub2>

Paper 4

Cotterill S, Powell R, Rhodes S, Brown B, Roberts J, Tang MY, Wilkinson J. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. 2019;8:176. <https://doi.org/10.1186/s13643-019-1077-6>

Paper 5

Cotterill S, Tang MY, Powell R, Howarth E, McGowan L, Roberts J, Brown B, Rhodes S. Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis. *Health Serv Deliv Res* 2020;8:41. <https://doi.org/10.3310/hsdr08410>

Paper 6

Tang MY, Rhodes S, Powell R, McGowan L, Howarth E, Brown B, Cotterill S. How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis. *Implementation Science*. 2021;16(1):8. <https://doi.org/10.1186/s13012-020-01072-1>

Paper 7

Rhodes S, Dias S, Wilkinson J, Cotterill S. Synthesis of data from trials of interventions designed to change health behaviour; a case study. Authorea. 2021.

<https://doi.org/10.22541/au.163256037.77153698/v1>

ACKNOWLEDGEMENTS

Thank you to my children for both inspiring me to do this and for accepting the lost evenings and weekends as I worked on it. For a few years as a stay-at home Mum in Egypt, I discovered that my children did not realise that a mother could go to work or a drive a car. I hope that I have shown you that being a parent is not an obstacle to success.

Thank you to Sarah for being an amazing supervisor and mentor. Your patience, kind words and constructive feedback are what has got me to this point.

Thank you to all of my collaborators for having faith in my input, for your generosity of time and your attention to detail.

Last but not least, thank you to my husband and parents. For always allowing me the freedom to choose my own path. Thanks for your unwavering support and encouragement.

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

When conducting randomised controlled trials (RCTs) assessing the effect of introducing a new health service or policy, researchers face challenges over and above those of clinical trials of medicines: a) a policy or service is usually implemented at an organisational level and affects everyone at that site; b) once a health worker has been trained in a new way of service delivery, any service they provide to the control group is likely to be contaminated by their training; c) patients and health workers in the intervention group may have contact with people in the control group and pass on information about the intervention they have received.

Cluster randomised trials (CRTs) (2) offer a solution to these challenges. In a CRT, groups of participants (clusters) are randomised to the same arm. Typical clusters in health care are people at the same hospital, hospital ward, GP surgery, or geographic region.

1.2 THESIS STRUCTURE

I present seven published research articles as evidence of my significant contribution as an applied statistician to the statistical design, analysis and synthesis of CRTs and the impact this has had on the literature within the applied fields that I have worked in.

1.3 OVERVIEW OF CHAPTER 2; DESIGN AND ANALYSIS OF CRTs

I was lead statistician on the Organising Support for Carers of Stroke Survivors (OSCARSS) trial (3-5), responsible for the design and analysis of the study from its inception. I present two papers from this trial (3, 4) to demonstrate the methods I developed to minimise bias in a CRT of a complex intervention.

The aim of the OSCARSS CRT was to test the effect of offering a new carer support intervention (staff training and new processes), compared to the current provision. As the intervention was implemented at the cluster level, all carers in the intervention clusters were offered the intervention, without the necessity for consent (6). Recruitment of carers to provide data came after the intervention had already started, which put the trial at risk of differential recruitment (7). Methods to minimise bias included prescriptive recruitment

strategy, carefully worded trial materials, monitoring of recruitment, retention and demographics and contingency in the Statistical Analysis Plan (SAP).

1.4 OVERVIEW OF CHAPTER 3; NOVEL METHODS TO INCORPORATE CLUSTER

RANDOMISED CROSSOVER TRIALS IN META-ANALYSIS

I present a Cochrane systematic review (8) to display innovative methods that I developed to incorporate trials with non-standard cluster designs into systematic reviews. The review evaluated the effect of chlorohexidine bathing to prevent infection in the critically ill. Many of the trials in this review were cluster randomised crossover (CRXO) trials (9) where a hospital Intensive Care Unit (ICU) was assigned to treatment or control for a period and then crossed over. This is a suitable study design in the ICU context, where there is a high changeover of patients, implementation is quick and there is little carryover effect. A crossover trial has additional statistical power compared to a parallel design because it utilises comparisons within the same cluster, as well as the parallel comparisons between control and intervention arms. A properly reported CRXO trial (10) will provide summary statistics with standard errors that take both the parallel and within cluster comparison into account, while adjusting for clustering. Standard methods (11) to incorporate CRTs into meta-analysis use the ICC and only take the parallel comparison into account; this wastes information and leads to estimates that are imprecise and potentially biased (giving most weight to the smaller, less robust trials). I implemented a novel statistical approach that utilises the full information reported in the trials. I converted all the summary data to a common format (rate ratio) and estimated appropriately adjusted standard errors to allow meta-analysis using the inverse variance method (12). I conclude this chapter with a set of generalizable recommendations for meta-analysis planning to incorporate CRXO into systematic reviews.

1.5 OVERVIEW OF CHAPTER 4; NOVEL METHODS OF META-ANALYSIS TO SYNTHESIZE

DATA FROM TRIALS OF COMPLEX INTERVENTIONS WITH MIXED LEVELS OF CLUSTERING

I discuss the SOCIAL (13-15) systematic review, funded by NIHR Health Services and Delivery Research funding stream. SOCIAL was a large systematic review of social norms interventions to change healthcare professional behaviour, on which I led the meta-analysis. Trials in this review aimed to estimate the effect of implementing strategies to

change the behaviour of groups of healthcare professionals. The trials were heterogeneous in their design. The units of randomisation included the health care patient, health professional, team, clinic, hospital or district. The units of analysis were the patient, the health professional or some larger unit. Meta-analysis of the data needed to take into account clustering at multiple levels (16). An additional complexity was that the trials reported many different behavioural outcomes (e.g. prescribing, test-ordering, hand-washing) in binary, ordinal or continuous format. I developed a suitable way of converting them to a common primary outcome measure of 'compliance with desired behaviour'. I present three papers on the design and results of this systematic review and a methods paper (17) that I led comparing alternative statistical methods (18-20) to synthesise mixed outcomes from trials of healthcare professional behaviour with mixed levels of clustering.

CHAPTER 2 DESIGN AND ANALYSIS OF CRTs

2.1 CHAPTER OVERVIEW

In this chapter I will summarise the current literature on the design of CRTs and reflect upon how I applied this knowledge to the decisions I made when designing and analysing the OSCARSS CRT (described in Paper 1 and Paper 2, presented at end of thesis).

I will start by describing the pragmatic approach to sample size estimation that we chose to use in OSCARSS. Sample size calculations for CRTs are more complex than calculations for individually RCTs and must take into account the degree of clustering, and may be constrained by the number of available clusters and likely cluster size.

Selection bias is a cause for concern in a CRT when participants within a cluster are recruited after the cluster has already been allocated to an intervention. I review the evidence on selection bias in published CRTs and methods to address the problem. I then describe the approaches that we used in the OSCARSS trial to minimise the risk of selection bias.

Finally I describe alternative methods of analysis for CRTs and the methods used in OSCARSS. I discuss how results were interpreted and the impact this had.

2.2 PRAGMATIC APPROACH TO SAMPLE SIZE ESTIMATION

2.2.1 Overview of methods to calculate sample size in CRTs

When planning a RCT it is important to consider sample size in advance. The usual purpose of an RCT is to provide an unbiased estimate the effectiveness of a treatment or intervention; this is done by comparing the measurement of a particular outcome of interest across intervention and control groups. The difference in outcome between groups is known as the 'treatment effect'. A trial that is large enough will have sufficient power to allow the treatment effect to be estimated with sufficient precision to enable conclusions (21). A small trial with low power will have low precision, providing wide confidence intervals for the treatment effect, leading to results that are inconclusive and the need for further research. A trial that recruits more participants than necessary may cause excess participant burden, research waste, or unnecessary harm (22).

Before embarking on a sample size calculation, it is important to choose a primary outcome measure, and determine the magnitude of the improvement in outcome that would be deemed to be important by key stakeholders such as patients and clinicians; this is known as the 'target difference'. Ideally, there should be some evidence that this target difference is realistic for the intervention e.g. using historical data or data from a different setting. The DELTA 2 guidance (23) describes methods to determine a target difference. For an individually-randomised parallel-group trial this target difference is used alongside other information about the primary outcome measure (such as the estimated control group rate for a binary measure and the estimated control group mean and variance for a continuous measure) to calculate the sample size required to detect the target difference with sufficient statistical power. Note that information on statistical power needs to be combined with other information about recruitment rates, eligible participants and other practical considerations (24).

In a CRT, rather than individuals, groups of participants (clusters) are randomised to the same arm. Typical clusters in health care are people at the same hospital, hospital ward, GP surgery, or geographical area. For CRTs, on top of the considerations for a RCT, researchers additionally need to consider the number of clusters, the size of each cluster and the degree of clustering.

Determining the units to be classed as the 'cluster' is not always straightforward. Often a CRT assesses the effect of introducing a new health service or policy. When a policy or service is implemented at an organisational level and affects everyone at that site then this naturally becomes the cluster. When a group of participants are to be offered a group therapy the cluster unit may be determined by practicalities such as room space or geography or how many patients a therapist can manage. Cluster randomisation may be used to avoid contamination between participants in different arms of the trial, and therefore it is important that clusters are independent— e.g. by avoiding staff working across more than one cluster.

The choice of clustering unit may lead to restrictions in terms of the number of clusters and/or the number of participants in each cluster. The number of available units may be limited; if the cluster is 'hospital' then researchers may be limited by the number of hospitals within a country. An entire cluster can be recruited simultaneously (e.g. school classes) or participants are recruited over time (e.g. cancer patients requiring surgery). The number of participants per cluster is likely to be restricted by the number of available

participants, the ability to recruit them, and the length time it will take. The size the cluster may also vary – e.g. hospital wards of differing sizes.

The ICC is a measure of the variation between clusters compared to within clusters (25). Outcome data from individuals within the same cluster may be more similar (correlated) than data from individuals from different clusters. This clustering tends to lead to sample sizes that are larger than those for equivalent individually RCTs, with greater levels of clustering leading to larger sample sizes.

One approach to sample size for CRTs is the use of the variance inflation factor (VIF)(26) or design effect. Researchers estimate the sample size for an individually RCT, and multiply by the VIF to allow for clustering. The VIF will depend on the ICC as well as the cluster size (n), where $VIF = 1 + (n-1)ICC$. This does not require additional software, but assumes a fixed cluster size and does not encourage exploration of other parameters.

Hemming et al. (2) describe methods to design efficient CRTs. If the number of available clusters are fixed then there is a point at which increasing the number of participants within a cluster makes very little difference to the power, and they call this the ‘point of diminishing returns’. Power curves are suggested as practical aids to help researchers determine the point of diminishing returns for a target difference. The authors suggest that the number of clusters and cluster size should be determined simultaneously rather than independently by graphically exploring a range of scenarios. A web-based R-shiny application enables researchers to graphically explore the relationship between cluster size, number of clusters, ICC, power and sample size without specialist software (27).

The *clsampsi* package (28) allows users to specify the variance of the cluster size and allows more complex clustering structures, including where the cluster size and variability is different across study arms.

A common problem in sample size calculations for CRTs is that the ICC is not known. It may seem appealing to conduct preliminary pilot or feasibility study with an attempt to estimate the ICC, but simulations (29) show that most pilot studies will be too small to estimate the ICC with sufficient precision. Some empirical studies (30, 31) aim to provide information about ‘typical’ ICCs in different settings, although the ICC will vary, not only according to the population and level of clustering, but also by the type of outcome measure too. A database of trials with reported ICCs (32) enables researchers to identify studies in similar populations with similar outcome measures on which to base estimates.

2.2.2 Example of a pragmatic approach to sample size calculation; OSCARSS

When designing the OSCARSS trial, I adopted a pragmatic approach to sample size (3). Very little was known about the ICC or the number of carers that could be recruited. A range of likely scenarios were explored based on certain constraints. This was to ensure that the trial would be likely to have sufficient power to be able to detect a target treatment effect, if it were to exist.

The first constraint was the number of clusters. The Stroke association has 12 UK regions each split into a number of areas, with each area split into a number of services. While region, area or service initially seemed like useful clustering units, some staff worked across multiple services or even areas. Some services were small, seeing less than one new carer per month: recruitment would likely be too low to justify the additional training costs. By grouping together services served by the same staff, the research team identified 36 independent stroke service units that were willing to take part in the OSCARSS trial and were likely to see at least 5 carers per month; these independent stroke service units were chosen to be the clusters. Assuming that some service units were likely to drop out, we expected that outcome data would be obtained from 24 to 32 service units, giving 12 to 16 clusters per arm.

The second constraint was the number of carers per cluster, which was limited by the size of the service, the number of carers they saw per month and the time-length of the trial. At the planning stage there was only 1.5 years of funding available and plans were to follow up all participants for 6 months, which allowed for only 9 months of active recruitment. Very little historical data existed about the number of carers seen by the Stroke Association, as their main remit is to support the stroke survivors, with minimal record-keeping about their contact with the carers of stroke survivors. Using data provided by the Stroke Association on stroke survivors as well as discussions with some Stroke Association staff, the Principle Investigator (PI) estimated that each cluster could recruit 4-6 carers per month. Using these estimates and a 20% dropout rate, outcome data from an average of 30 to 45 carers per cluster was assumed for initial sample size calculations.

The primary outcome measure was caregiver burden, which was captured using a subscale of the Family Appraisal of Caregiving Questionnaire (FACQ) (34). This subscale consisted of 8 questions using a 5 point Likert response; with the mean score per question calculated. Cooper (34) reported mean (standard deviation (SD)) of 3.13 (0.87) for this scale for carers in a palliative care setting. I had discussions with other researchers in the team to decide

what the target difference on this scale would be. Members of the study Research User Group (RUG) were also consulted by one of the PIs. The 8 question scale could either be summarised using a total score out of 40 or a mean score out of 5. It was felt that a reduction would need to be at least 2 or 3 points on the total score to be a meaningful improvement for a carer. A 2 point difference could mean going down from strongly agree to neutral on one question, or from agree to neutral on two questions, or equivalent improvement. This corresponds to a 0.25 difference in the mean score. A 3 point difference would, for example, mean changing from agree to neutral on three questions, which corresponds to a 0.375 difference in the mean score. Any of these differences would suggest on average a real improvement in at least one aspect of carer burden for each carer; discussion with the RUG suggested that anything less than this seemed too small to be meaningful. In a stepped-wedge trial of the same carer support intervention within palliative care (35) the same support approach led to a mean improvement of 0.31 points on the same outcome measure, so the chosen effect sizes were considered by the team to be both meaningful and realistic. As a pragmatic approach, power was calculated for a range of effect sizes, including a 0.25 point, 0.31 and 0.375 point improvement in the mean score.

Similar studies in the database of ICCs in implementation trials (32) reported ICCs which ranged from 0.01 to 0.05. Based on this, I felt it likely that the ICC would be no higher than 0.05, but I performed calculations for a range of values from 0.01 to 0.1 to include best and worst case scenarios. The TRACS CRT (33), which focussed on carers for stroke survivors reported that the ICC was 0.027 so this provided reassurance that our estimates for the ICC in OSCARSS were of the right order of magnitude.

Table 2-1 shows a selection of the scenarios that I explored while designing the OSCARSS trial, to show power would vary according to effect size, cluster size, number of clusters and ICC. I presented information and led a discussion with the Trial Management Group (TMG) before the trial started. I demonstrated that the trial appeared to have power to detect a 0.375 point difference or more under a range of plausible parameters, including if we dropped to as few as only 12 clusters per arm or had an ICC as high as 0.1. I also showed that the trial was unlikely to have sufficient power to detect a 0.25 point difference unless we had at least 16 clusters per arm and the ICC was no more than 0.05. I used the Stata *clsampsi* command.

Table 2-1 Power projections (assuming SD = 0.9)

ICC	Power to detect given detect size			
	0.375 point reduction in primary outcome		0.25 point reduction in primary outcome	
	12 clusters per arm 45 per cluster	12 clusters per arm 30 per cluster	12 clusters per arm 45 per cluster	12 clusters per arm 30 per cluster
0.01	100%	100%	95.2%	87.8%
0.05	95%	93%	68.1%	62.0%
0.075	89%	85%	55.4%	51.2%
0.1	77%	80%	46.4%	43.5%
	16 clusters per arm 45 per cluster	16 clusters per arm 30 per cluster	16 clusters per arm 45 per cluster	16 clusters per arm 30 per cluster
0.01	100%	100%	99%	96%
0.05	99.0%	98.0%	81%	76%
0.075	95.8%	93.8%	69%	65%
0.1	90.8%	88.4%	59%	56%

Table 2-2 shows how the total required sample size would vary according to the ICC and the anticipated between-group difference in primary outcome. I presented data to the TMG in a variety of different ways to illustrate why there was uncertainty in our sample size estimates and how different ICCs and recruitment scenarios would affect power. Having worked in individually RCTs, some of the research team were keen to aim for a fixed target sample size. I stressed to the team that while the total sample size is important, it is also important that all clusters contribute outcome data rather than all the data coming from a small number of high recruiting clusters. I suggested that the research team modify how recruitment data were presented in reports for the TMG and the Trial Steering Group (TSG); we produced tables with both the overall totals and the recruitment per cluster which allowed us to target clusters with little or no recruitment for support. I also stressed that there is uncertainty in our estimates because we didn't know what the true ICC was going to be, and therefore we should recruit as many participants as possible in the available time frame in case the ICC was higher than expected.

As described in our protocol (3), we aimed for a minimum target of 320 carers from at least 32 clusters of roughly equal size providing primary outcomes at three months. This would allow us 80% power to detect effect sizes of 0.31 or more for ICCs ≤ 0.01 , and effect sizes of ≥ 0.375 for ICCs of ≤ 0.05 . We assumed a retention rate of 80% between consent and primary outcomes, which meant that we required a minimum of 400 consented carers. A recent review of RCTs funded by the NIHR HTA programme (36) found that over 75% of trials had a retention rate of 79% or more suggesting that 80% is a reasonable estimate of retention in the absence of more relevant data. We also quoted that our 'optimal' sample size was outcome data from 512 carers and planned that we would only stop recruitment early if we reached this total; this would prevent us from potentially missing more subtle target differences of 0.25 if we had the resources to do this.

Table 2-2 Total sample sizes to achieve 80% power (assuming SD=0.9 and 16 clusters per arm)

ICC	Effect size		
	0.375 (3 points)	0.31 (Aoun et al.)	0.25 2 points
0.01	224	320	512
0.05	288	512	1312
0.075	352	800	
0.1	380	988	

During the trial recruitment was monitored closely, and it became apparent that the clusters were recruiting at different rates leading to large differences in cluster size, with apparently more variability in the intervention arm than the control arm.

The coefficient of variation (COV) of the cluster size is defined as the ratio of the standard deviation of cluster size to the mean cluster size (37, 38). This can be included in sample size calculations using the *clustersamps* command (39) to allow for varying cluster sizes.

Ten months into the 18 month trial, in order to estimate the standard deviation of the cluster size, I produced Table 2-3 which shows actual values at 10 months and projections for 18 months of the mean (SD) cluster size in OSCARSS. I used the 10 month data and assumed that each cluster would continue to recruit at a constant rate until the 18th month with 20% dropout to produce a crude estimate of the mean and SD of the final cluster sizes, leading to estimates of the COV of 0.99 and 0.68; I have used the larger of the two to investigate the effect of variation up to this magnitude on power..

Table 2-4 shows how the cluster size variation would impact power for a variety of effect sizes and ICCs as before using the estimated coefficient of variation of 0.99. Comparing power for a fixed cluster size compared to the predicted scenario, it seemed that having varied cluster sizes would reduce power, with the greatest impact seen when the ICC is high. By this point, we were already monitoring the cluster sizes throughout the trial, giving extra support and encouragement to low recruiting centres to try to reduce cluster size variability.

Table 2-3 Projected variation in cluster size after 10 months

	Intervention	Control
Currently Recruited		
Number of Clusters	18	17
Cluster size (Mean)	7.8	7.1
SD	7.8	4.8
Projected total recruitment by end of trial*		
Clusters	18	17
Mean	11.3	10.2
SD	11.2	6.9
Estimated coefficient of variation	0.99	0.68

*Assuming 20% drop out and 18/10 of current recruitment based on 10 months of recruitment so far and 8 months remaining

Table 2-4 Projected impact of variation in cluster size (assuming 18 clusters, mean cluster sizes of 10 carers per cluster, SD=0.9)

Mean difference between intervention and control groups and ICC	Expected power to detect difference	
	Assuming no variability between arms in cluster size (COV=0)	Assuming COV= 0.99
Difference=0.375		
ICC=0.01	96%	94%
ICC=0.025	93%	89%
ICC=0.05	89%	79%
ICC=0.075	84%	70%
Difference=0.31		
ICC=0.01	86%	83%
ICC=0.025	82%	75%
ICC=0.05	75%	63%
ICC=0.075	69%	53%
Difference=0.25		
ICC=0.01	69%	65%
ICC=0.025	64%	56%
ICC=0.05	57%	45%
ICC=0.075	51%	38%

During the first 9 months of recruitment it became apparent that the initial rates of estimated recruitment were over optimistic; we were allowed a time extension to recruit to target over a period of 18 months. However, other estimated parameters proved to be reasonable accurate. Between 1 February 2017 and 31 July 2018 a total of 414 carers

were recruited from 35 randomised clusters (18 intervention; 17 control). In line with our estimated retention rate of 80%, 84% of recruited participants provided outcome data (175 intervention; 174 control). The mean (SD) FACQ carer strain at 3 months was 3.11 (0.87) in the control group compared with 3.03 (0.90) in the intervention group, adjusted mean difference of -0.04 (95% CI -0.20 to 0.13) (4), so the observed standard deviations were very close to the 0.9 used in the sample size calculations. The ICC for the primary outcome measure was 0.02, similar to the previous TRACS trial in stroke carers (33). The tight confidence interval around the effect estimate rules out any meaningful difference in average outcome. This suggests that the trial was sufficiently powered, and we could conclude that the CSNAT Stroke intervention, as implemented in OSCARSS did not improve carer burden when compared to the usual level support.

2.3 METHODS TO MINIMISE SELECTION BIAS

2.3.1 Evidence of selection bias in CRTs

In RCTs, randomisation is used to ensure that participants in each arm come from the same population. Selection bias occurs when participants selectively enter the trial (or not) based on knowledge of what their treatment allocation is likely to be (40). This leads to trial arms that are no longer representative of the same pool of participants. To avoid selection bias, trials can adopt steps to maintain allocation concealment (41), ensuring that, at the point of recruitment to a trial, neither the potential participant nor the recruiter has knowledge of the next treatment allocation.

In a CRT, it is the clusters that are randomised rather than individuals. The clusters are often randomised simultaneously in a single step at the beginning of the trial, with the intervention implemented at the level of the cluster. Where routinely collected data is utilised for the outcome data there may not be a need to recruit individual participants. For example in a trial of a hospital wide strategy to reduce MRSA infection (42); in this situation the hospital rate of MRSA infection is routinely collected and reported so there is no need to identify or recruit individual patients and the risk of selection bias is low. Where individual participant consent is required, for ethical reasons and/or because of additional data collection, this recruitment is likely to occur after the cluster allocation has been revealed to staff or researchers working within the cluster, and therefore the risk of selection bias may be high.

Selection bias may be evident as differential recruitment which can mean both a difference between the rate of recruitment across study arms and/or differences in participant characteristics (43). Note also that selection bias could occur at the cluster level if allocation is revealed to clusters before they formally agree to take part. A situation when a trial has 'empty clusters' (44) because a cluster agrees to take part and then chooses not to recruit once they become aware of their allocation is another form of selection bias.

A CRT (45) compared care by a centralized clinical pharmacist to usual care. The 'cluster' in this case was a rural primary care office, and staff and patients were aware of the allocation of the office to either intervention or control arm at the time of recruitment. It was observed that patients with poorly controlled diabetes were less likely to consent in the intervention sites compared to control. In addition, the staff at the control sites may have recruited more complicated patients. In a second example of differential recruitment (7), clusters were randomised to either receive training in active management or continue with their usual care. On average practices in the active management arm recruited 12.7 participants, while practices in the control arm recruited only 5.1 participants. Participants recruited by practices in the active management arm tended to be more likely to be working full-time, more highly educated, and have less symptoms. The intervention in this case included a training element (43) where practice staff were educated in diagnosis and therefore the intervention itself in this case is highly likely to have contributed to the selection bias, in addition to the awareness of the intervention. The evidence of selection bias led the researchers to revise their trial design before the full trial.

A review of 36 CRTs in key journals (46) found evidence of differential recruitment in seven (30%) of the 23 trials where participants were selected after randomisation. Only 21 out of 34 (62%) CRTs in a primary care setting described methods that protected them against bias when recruiting patients (47). Among 24 CRTs published in leading medical journals 8 used methods of recruitment which left the trial at risk of selection bias, of these 5 trials (63%) had evidence of differential recruitment(48). Comparisons between individual RCTs and CRTs (43, 49) provide further empirical evidence of selection bias in CRTs.

2.3.2 Using modified informed consent to minimise selection bias

Selection bias is a risk in a CRTs when recruitment occurs after the cluster allocation has been revealed to participants, researchers or other personnel. While the default position is that informed consent is a key ethical and legal requirement for RCTs (50) and CRTs, there

are circumstances where, with approval by an appropriate ethics body, the approach to informed consent may be modified. Any modifications to the default position would necessarily be justified by very clear practical or scientific purpose(51).

Informed consent requires the participant to fully understand the aims and activities within the trial and voluntarily agree to the procedures which will include randomisation, intervention and data collection. When it comes to CRTs the requirements for informed consent are complex (52). In some cases it may be necessary for the entire cluster to be randomised and given the intervention before an individual participant is identified and without their knowledge e.g. health promotion poster campaign targeted at patients attending GP surgeries; here it would be impossible to take informed consent prior to randomisation or exposure to the intervention. Where the intervention is introduced at a hospital level (e.g. encouraging more frequent hand washing amongst staff to reduce infections) it may be impractical to take individual informed consent from every patient. It is important to consider ethical issues such as risks, autonomy, justice and respect alongside the scientific benefits of the trial when considering the need for informed consent (53, 54).

In an RCT, the participant is usually consenting to randomisation, the delivery of the intervention and data collection; in a CRT each of these should be considered separately (55). Note that these elements may impact different participants in different ways (56), so it is important to consider who the research participants are and which parts of the trial are relevant to them – e.g. an intervention may target health care professionals but data collection may be needed from patients.

Consent may need to be considered at both the level of the cluster and the level of the individual participant. When an intervention is delivered at a group level, it is often unclear whether consent is required for individual participants or a group representative(55). The CRT literature often refers to ‘gatekeepers’ with authority to give consent for some aspects of the trial on behalf of the cluster (57). For example, a head teacher (gatekeeper) may agree for a school to be randomised and for staff to have training in a new learning activity, while parents and children may agree to participate in the activity and provide data.

Guidance was developed in an attempt to provide guidance to researchers and research ethics committees (RECs) about the how to ethically conduct CRTs (58). Table 2-5 lists the items that relate to recruitment of individuals and clusters.

Recent suggested refinements (59) (56) add that ‘An REC may approve a modified consent procedure if there is a risk of contamination bias. This procedure implies that randomization should not be disclosed under certain conditions.’ puts emphasis on the need to minimise bias during consent procedures. Although it is contamination bias that is mentioned here, a modified consent procedure can also help prevent selection bias. A modified procedure could allow consent when participants are not fully informed about every element of the trial; for example they may be informed about data collection procedures but not about randomisation or the aims of the trial. If a participant is unaware that they are part of a trial, and do not realise that have been allocated to an intervention that is anything other than standard care, then the decision to take part will likely be based on the burden of participation rather than any expectation about benefit from an intervention, hence minimising selection bias. Where participants have been involved in research without a full informed consent procedure it is good practice to debrief them afterwards (60).

Table 2-5 Items from the Ottawa Statement on the Ethical Design and Conduct of CRTs relating to recruitment

Ethical issue	Item number	Recommendation
Obtaining informed consent	4	Researchers must obtain informed consent from research participants unless a waiver of consent is granted by a REC under specific circumstances.
	5	When participants’ informed consent is required, but recruitment of participants is not possible before randomisation of clusters, researchers must seek participants’ consent for trial enrolment as soon as possible after cluster randomisation—that is, as soon as the potential participant has been identified, but before the participant has undergone any study interventions or data collection procedures.

	6	A REC may approve a waiver or alteration of consent requirements when the research is not feasible without a waiver or alteration of consent, and the study interventions and data collection procedures pose no more than minimal risk
	7	Researchers must obtain informed consent from professionals or other service providers who are research participants unless conditions for a waiver or alteration of consent are met
Gatekeepers	8	Gatekeepers should not provide proxy consent on behalf of individuals in their cluster
	9	When a CRT may substantially affect cluster or organisational interests, and a gatekeeper possesses the legitimate authority to make decisions on the cluster or organisation's behalf, the researcher should obtain the gatekeeper's permission to enrol the cluster or organisation in the trial. Such permission does not replace the need for the informed consent of research participants
	10	When CRT interventions may substantially affect cluster interests, researchers should seek to protect cluster interests through cluster consultation to inform study design, conduct, and reporting. Where relevant, gatekeepers can often facilitate such a consultation

A different modified approach to informed consent could be an 'opt-out' procedure (6). There are situations where it may be physically impossible for a participant to refrain from exposure to the intervention (e.g. a leaflet campaign encouraging mask wearing to prevent SARS-CoV-2 infections) but if they are informed about the trial they could be given the opportunity to opt-out of allowing their data to be included.

2.3.3 Using recruitment before randomisation to minimise selection bias

An alternative method to avoid selection bias is selecting and recruiting participants before the clusters are randomised. This method is possible in cases where entire clusters can be recruited simultaneously and followed over time. An example of this is where a school class is the cluster (61); the entire class could be given information on the study, randomisation and the potential interventions and asked to give consent before randomisation. This approach can lead to long delays between recruitment and the implementation of the intervention(61). In a survey of 113 authors of reports of CRTs (62) 44 (39%) reported that they recruited participants before the randomisation of clusters (although this was only documented in 9 trial reports).

2.3.4 Other methods to minimise selection bias

Another suggestion (43) is to use an independent person to undertake recruitment. In order to prevent selection bias this person should be masked from knowledge of the intervention allocation (44). To avoid empty clusters, a cluster should not be randomised until after the first participant has been recruited (22), although this could be difficult if the intervention involves an element of staff training or the introduction of new equipment. A survey of trialists (62) found that 8 (7%) reported that they had used a blinded recruiter (although this was only evident in 1 trial report).

2.3.5 Methods used in OSCARSS to minimise selection bias

In the OSCARSS trial participants within clusters were the carers of Stroke survivors who were newly receiving support by the Stroke Association. The intervention was aimed at new carers so it was impossible to identify and recruit carers before the clusters were randomised; instead new carers were identified and supported by Stroke Association staff as they presented to the service. No routine data are collected about carers of stroke survivors, so we needed to ask participants for both demographic data and outcome data. The trial introduced several major challenges in recruiting carers, while minimising

selection bias. Selection bias could occur both during the selection of eligible carers by staff and during the decision of the individual carer about whether or not to take part.

The intervention, the Carer Support Needs Assessment Tool (CSNAT) approach (4) involved training for staff in i) the identification of carers; ii) provision of suitable carer support; iii) methods to engage with the carers, which included arranging a one-to-one face-to-face meeting, in contrast to the standard support where the carers were not seen separately to the cared-for stroke survivor. Originally, members of the OSCARSS team wanted to use different recruitment methods in the two arms of the trial with carers in the control arm contacted over the phone while carers in the intervention arm would be introduced to the trial during their face-to-face meeting. I was very concerned that different methods of participant selection would lead to bias.

Recruitment was a two stage process with stroke service staff initially approaching the carer, providing information about the study and then asking them whether they'd be willing to be contacted by researchers at the University of Manchester. The second stage involved a phone call from the OSCARSS research team. I was concerned that the recruitment process could not be blind as it was initiated by the same stroke staff who were delivering the service. As a team, we became worried that staff would deliberately or inadvertently 'cherry pick' participants who were most likely to be positive about the support they received, and that this could vary by arm e.g. if control arm staff were disappointed not to receive the intervention. We were looking for the effect of the 'implementation' of the new approach to the entire cluster so I was also keen to stress that we wanted data from all carers wherever possible, not only those who engaged with the intervention i.e. an intention-to-treat approach (63). The second stage of recruitment was also not blind as the staff doing the recruitment were the same staff that filled in the database and contacted cluster staff to collect information; with extra resources this process could have been carried out blind by an additional researcher.

I stressed to the team the importance of a method of recruitment that (a) was identical in both arms and (b) independent of engagement with either the control or experimental intervention. To address this, as a team we developed training for Stroke Association staff in both arms of the study with very prescribed details about the identification and recruitment of carers. The staff were informed about the design of the study and its aims and the need to minimise bias by identifying and approaching all carers, regardless of whether or not they interacted with the intervention materials and regardless of the stage

in the stroke journey (note that a stroke-survivor and their carer may approach the Stroke Association for support weeks, months or even years after the time of the stroke). In both arms, staff were encouraged to seek out the carer when visiting the stroke survivor. Study information materials were identical in both arms. Our RUG were involved in the design of staff training and participant materials. They were keen to make it clear that the opportunity to take part in research was the right of each eligible carer and that the stroke association staff should approach all carers without making assumptions on their behalf about the burden of taking part.

I suggested that a modified approach to informed consent be adopted. The intervention consists of a package of staff training, new paperwork as well as a person-centred approach to carer support; many aspects of it are delivered at a cluster level and start the moment a carer interacts with the service; for these it would be unfeasible to ask for individual consent. For other elements, such as the use of the CSNAT tool to discuss carer support needs, participants had the freedom to engage as they wish. I encouraged the team and our RUG to weigh up the harms and benefits of asking participants about the randomisation process and asking for consent midway through an intervention. I suggested that the participants be asked to provide informed consent for being part of a research study and to having their data collected, but without informing them that they were taking part in a RCT. We designed participant materials that told participants that they were part of a research study assessing the quality of support for carers. The RUG were very supportive of this approach and collectively we felt that there was minimal risk of harm from the intervention, and for the parts that were delivered at an individual level, participants were implying their consent (or not) by their level of engagement. Participants were being fully informed about the study aims and the burden of taking part before allowing the use of their data. The REC approved the study without query. Once we had adopted the modified consent procedure, I pointed out to members of the team that we had to avoid any reference to the fact that OSCARSS was a trial in all publicity, this was so as not to confuse carers or lead them to feel they had been deceived. After the trial, the trial team produced summary materials for trial participants and the stroke community that I helped to edit (<https://arc-gm.nihr.ac.uk/projects/oscarss>).

Despite the steps taken to minimise the selection bias we still felt our trial was at risk. The PIs and I felt there was a need to address this risk in our plans for analysis. In the SAP I included a section on differential recruitment which included the design features described above. I also decided that our primary analysis would be adjusted for baseline covariates, in

an attempt to address any minor baseline imbalance (44). As pre-specified in the protocol we adjusted for individual level covariates: time post-stroke; age of carer; health of carer at study entry; stroke severity (as rated by carer); and the cluster level covariates: size of service and experience of staff delivering support. These variables were chosen by the PIs as being likely to be strongly associated with the outcome. Size of service was used in stratified randomisation of the clusters. I also included a statement about what we would do if we or our independent Trial Steering Group (TSG) had serious concerns about differential recruitment. I wrote that in the event of serious differential recruitment we would seek independent advice from our TSC and either (a) analyse as observational data using propensity score methods (64) (65) or (b) not analyse outcome data. Note that while examples of the use of propensity scores to analyse cluster trials can be found (66) their use for continuous outcomes was not well supported by a simulation study (65) and if the TSG had recommended this type of analysis, conclusions would have been very cautious indeed.

Once the trial had started I monitored it throughout to check for issues with selection bias. 'In reports to the TMG (held every 2 months) and the TSG (held every 6 months) we produced tables of both rates of recruitment and retention (by cluster by arm) and carer and stroke survivor demographics (by arm only without adjustment for clustering) including age, relationship with stroke survivor, time since stroke, gender'. The chair of the TSG was a senior CTU statistician with extensive cluster trials experience. We highlighted the need to look for differential recruitment but did not provide specific guidance. At one point it did look as though the average time since stroke was longer in the intervention arm and this led to the Trial Manager contacting staff in both arms, reminding them of the eligibility criteria and the need to approach all new carers. However no imbalance was observed in either rate of recruitment or the carer characteristics by the end of the trial (4).

2.4 ANALYSIS AND INTERPRETATION

2.4.1 Overview of methods of analysis for CRTs

When analysing individually RCTs the method of analysis is likely to be dictated by the type of outcome measure (e.g. binary, continuous, survival), the number of arms, any repeated measurements over time (67) and any adjustment for baseline covariates (68) . Typical analysis methods are simple hypothesis tests (e.g. t-test), and regression methods (e.g. ANOVA, logistic regression) which assume independence amongst data points. When

analysing a CRT, it is important also to consider clustering at the level of randomisation. Ignoring the clustering would (i) violate independence assumptions and (ii) lead to overly narrow confidence intervals.

One method of analysis for CRTs is to simply aggregate data at a cluster level before analysis (69). Each cluster provides a single summary measure, for example the mean over all members of the cluster or the proportion of cluster members with a binary outcome. As each cluster provides a single data point, data are independent and standard techniques can be used. If the clusters vary in size it may be desirable to weight the analysis by cluster size (70). Regression techniques may be used to adjust for baseline covariates if these are measured at a cluster level. This type of analysis does not require specialist software or knowledge, but it is likely to result in less statistical power than other methods (69).

An alternative method is to conduct standard analyses at the level of the participant and then make adjustments using the design effect to take clustering into account. For example, a correction factor can be applied to the test statistic (71) or the standard error (72). As with the aggregate method, this requires only basic statistical software and application of simple formulae.

Random effects models (73) are an extension of regression models with the addition of an extra random error term. These models are part of a family of models that may also be known as mixed effects models, multilevel models or hierarchical models. These conditional models take into account correlation between measurements from individuals within the same cluster via a cluster specific error term $u_j \sim N(0, \tau^2)$, where u_j is the departure from the group mean for cluster j . In this case 'cluster' would become what is known as a 'random effect'. This is in addition to the individual level error term $e_{ij} \sim N(0, \sigma^2)$, where e_{ij} is the error (departure from the cluster level mean) for person i within cluster j . When analysing a CRT, the treatment effect is estimated using a fixed term in the model. Baseline covariates can also be included in the model as fixed effects. This type of model can be extended to allow analysis of repeated measurements from the same individual over time by the addition of a third random error term for individuals, nested within clusters (74). Statistical software (Stata, R, SPSS) enables analysis using random effects models.

Generalised estimating equations (GEE) are marginal (population average) models that assume that the variance is a function of the mean. Correlation structures are defined for observations within the same cluster, and GEEs are thought to be reliable even when these

are miss-specified (75). It is common in CRTs to choose the 'exchangeable' correlation matrix with $\text{Corr}(Y_{ij}, Y_{kj}) = \alpha$ when $i \neq k$, where observation Y_{ij} is the outcome measurement for the i^{th} participant in the j^{th} cluster. As with random effects models, baseline covariates can be included and statistical software is required.

When choosing an appropriate method of analysis, the number of clusters is an important consideration. A simulation study (76) has shown that CRTs with a low number of clusters (70 or less) analysed by multilevel models or GEE are at risk of higher than expected rate of false significant results. Similar studies (75, 77, 78) found that neither type of model performs well with a small number of clusters and large ICC. Correction factors to reduce the inflated type one error are recommended (79).

An additional consideration when choosing a method is power. A comparison of methods to analyse CRTs with a binary outcome (80) showed with simulation that the GEE approach generally has the highest power but that the difference between that and the random-effects method is negligible. Similarly, other studies found very similar effect estimates from both GEE and random-effects models. .

One 2016 review of 96 CRTs (81) found that 4(5%) used aggregated data at a cluster level, 22(26%) analysed data at an individual level using basic statistical tests or regression models (with or without adjustment for clustering), 14(16%) used GEE and 45(52%) used a random effects model, with at least 74% using a valid method that adjusts for clustering. A second (76), also conducted in 2016, found that at least 86% of a random sample of 100 CRTs adjusted for clustering, although 65% were found not to have adequately taken into account the risk of inflated type-1 error due to a small number of clusters. An earlier review published in 1995 (82) found that only 12(57%) took clustering into account, suggesting that methods have improved over time.

2.4.2 OSCARSS analysis and interpretation

For the primary analysis of OSCARSS data I chose to use a random effects model. As seen above, this type of analysis is commonly seen in the CRTs literature and I was already familiar with this family of models and the Stata packages required. I was also interested in being able to explore the trajectory of the primary outcome over time, and I know that a random effects model would allow me to do this, even if there were not complete outcome data available for each individual at each time point.

At the time OSCARSS was conducted I did not consider the potential for an inflated type 1 error due to a small number of clusters. Eldridge et al. (83) recommend a minimum of 30 clusters when using parametric methods based on the normal distribution and with 35 clusters OSCARSS is large enough by this criteria. Kahan et al. (76) suggest that trials with less than 70 clusters could be at risk of inflated type 1 errors. Retrospectively, given that no evidence of an effect was observed (mean difference -0.04 (95% CI -0.20 to 0.13)), this is not a concern for OSCARSS but is important to take into account in future trials.

As lead statistician on the trial, I was supervising the OSCARSS analysis rather than carrying out the analysis myself, so it was important to specify in the protocol (3) and SAP (84) exactly how I expected the analysis to be carried out, as well as checking each stage against our pre-specification. While the OSCARSS trial was not run under the auspices of a clinical trials unit (CTU), I was keen for the trial analysis to be carried out to the same standard as trials within our local CTU. Several things that I introduced after reading standard operating procedures used within the CTU were (a) independent peer review of the SAP (b) independent programming of the primary outcome measurement and (c) blind preparation of the dataset using dummy codes for the group variable. In addition I was influential in making sure that our protocol and SAP were published before our analysis began. These measures ensured that the plans for analysis were clear and valid, and provided auditable proof that we followed our pre-specified plans without deviation. Measure (c) allowed any post-hoc decisions around data cleaning and data manipulation to be carried out without knowledge of group assignment and therefore avoid any unconscious bias that could influence the estimate of the treatment effect.

The analysis, which used a multilevel model approach via the *xtmixed* command in Stata (85) was reported in detail in the protocol with accompanying SAP (3) and trial report (4). As reported in section 2.2.1 there was no evidence of a difference between arms in terms of our primary outcome. In terms of interpretation of these results, in addition the magnitude of intervention effect and 95% CI I encouraged the team to look at the summary data by group. We utilised the manuals of the outcome measures to identify that the observed mean FACQ carer strain score of 3 in both arms indicated a neutral score (34) on average; perhaps an indication of adequate support in both arms. I created box and whisker plots **Error! Reference source not found.** and dot plots Figure 2-2 to display the full range of the data and show that while carer strain was fairly low on average, there still were a number of carers in each arm experiencing considerable levels of strain; these plots were used in posters and conference presentations by the PIs.

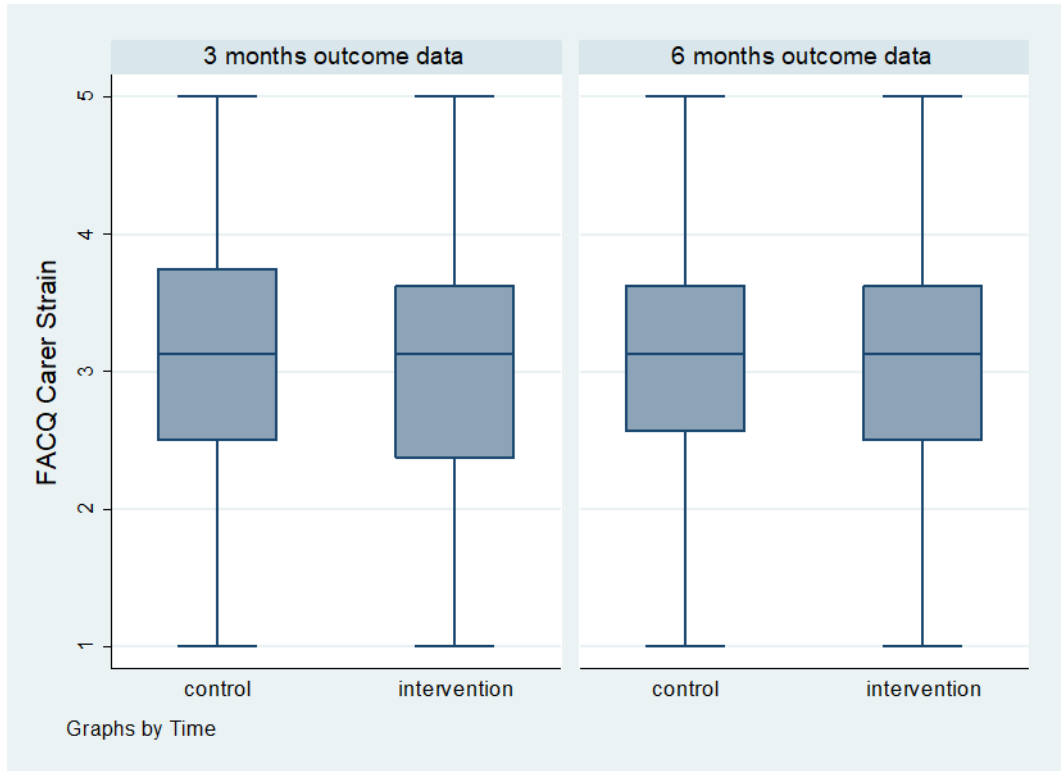
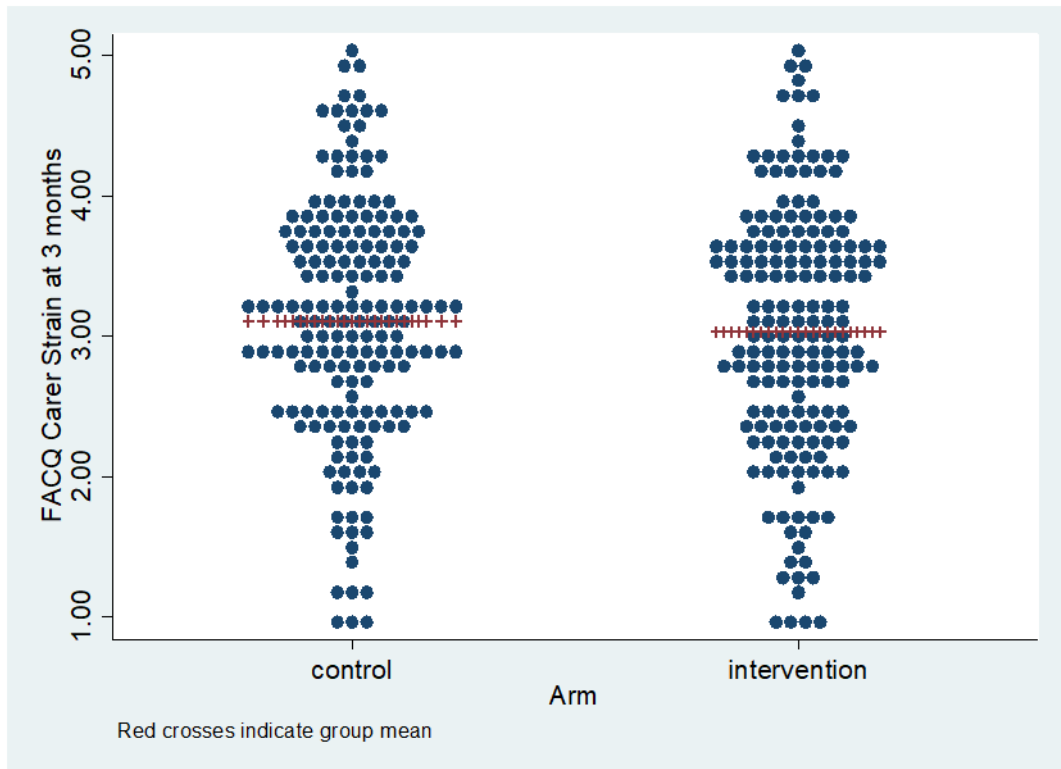


Figure 2-1 Box and whisker plot of OSCARSS primary outcome measure at 3 and 6 months, by group. Horizontal lines indicate 75th percentile, median and 25th percentile.



2.5 REFLECTIONS

I was involved in OSCARSS from the very beginning when I discussed with the PIs the pros and cons of different designs and persuaded them to undertake a CRT rather than a simple before and after study or stepped wedge design. We had limited time and resources, but I felt that if we were going to undertake any research at all then it should be as robust a design as possible. In OSCARSS we managed to run a large parallel group CRT without external funding or a CTU, relying on the existing staff and infrastructure available in the collaborating organisations.

During this project I learnt from others about leadership, management, organisation, recruitment, public involvement and communication. OSCARSS gave me the opportunity to be lead statistician on a large trial, to have a role in patient involvement, and to present at a CRT methodology conference. OSCARSS brought together a team of researchers who were passionate about research, good at listening to others, interested in the detail, and effective at communicating. It meant long discussions every step of the way, but also meant a great working relationship where my ideas were respected and I felt comfortable taking the lead statistically, expressing concerns, bouncing ideas off people, and trying out new things. One of the PIs said 'how amazing you were at communicating things to the reps of the RUG who sat on the TMG meetings..... I think you were really good at doing public involvement within the context of your role and you were also really good at collaborating with us stats dummies about rationale for things and – to me, that is gold and is a pivotal practical issue as a statistician on CRTs that will tend to involve a wide range of folks' so it is good to know that my efforts to communicate well did not go unnoticed.

Writing this thesis has caused me to reflect more deeply on issues of informed consent in cluster randomised trials. I believe that the approach we took in OSCARSS was proportionate to the harms involved and respected ethical principles; however in future cluster trials I would encourage the team to consider the balance of scientific integrity versus autonomy more thoroughly before deviating from the default position of fully informed consent for all active participants. I have encountered trialists that see cluster randomised trials as a simple way around the need for informed consent and I now feel more fully equipped to navigate the issues.

The results of OSCARSS demonstrated that the CSNAT carer intervention did not improve the burden for carers by any meaningful degree. Our process evaluation (5) revealed that the intervention was not implemented as intended. Had we run a small feasibility trial (86) first we may have rectified the issues with intervention fidelity or have abandoned plans for a trial. I have learnt from this the potential advantage of running a feasibility study in advance of a full trial to prevent research waste and reduce participant burden, although the benefits of this need to be weighed up against delays to the adoption of effective treatments. I also now understand more about the value of qualitative research in assessing intervention fidelity and implementation.

One thing I may have done differently if I were to design another CRT is to consider a baseline period (87). It has been shown recently that in some situations (88) a period of baseline data collection in advance of the introduction of the intervention can provide some additional power and reduce the overall sample size required, but I did not learn about this work until part way through the trial. This design has some of the advantage of a stepped wedge trial (89) in that it combines both within and between cluster assessments of intervention effects. In summary, I was instrumental in the design of OSCARSS, and implemented a number of steps to minimise selection bias. I also successfully estimated and explored the parameters required in a sample size calculation to ensure that the trial was very unlikely to be underpowered. I planned and managed an analysis strategy that was predefined, appropriate and rigorous. These elements together guaranteed that we had a trial that was robust with a result that we could believe.

2.6 METRICS

Paper 1 was published in *Trials* which has a 5 year impact factor of 2.6. It focussed on the performance and findings of trials in healthcare and has editors that are well respected in the realm of trials methodology. Importantly, this journal encourages publication of protocols and SAPs. This paper has over 1800 accesses and 4 citations.

Paper 2 was published in *BMJ Open*, which is dedicated to publishing medical research from any discipline and has an impact factor of 2.7. This paper has had over 2000 full text downloads and 3 citations. It has an Altmetric Attention score of 14 which puts it in the top 10% of research outputs tracked by Altmetric.

Two citing articles relate to planned trials developing alternative tools to support carers of stroke survivors (90, 91) and utilise lessons learnt from OSCARSS. Work on the OSCARSS trial has promoted the importance of research about carers, has prevented an intervention that was ineffective as implemented from further investment of precious resources, and has provided baseline information about carers for future research into carer support.

As a result of my experience working on OSCARSS I've been appointed onto a trial steering committee of a CRT by the NIHR Public Health Research Programme.

CHAPTER 3 NOVEL METHODS TO INCORPORATE CLUSTER RANDOMISED CROSSOVER TRIALS IN META-ANALYSIS

3.1 CHAPTER OVERVIEW

In chapter 2 I wrote about the design and analysis of CRTs. In this chapter I move onto situations where the researcher has access to reported data on multiple trials, including CRTs, and wishes to combine them using meta-analysis. I focus on a specific type of CRT, the cluster randomised crossover (CRXO) trial, as this is what I encountered in a systematic review I was working on, reported in Paper 3. The Cochrane Handbook (72) provides guidance on how to incorporate cluster randomised trials with imperfect reporting into meta-analyses but very little guidance exists on how to incorporate CRXO designs into meta-analysis. Here I describe methods I applied myself and use these to derive recommendations for researchers intending to include CRXO trials in meta-analyses. .

3.2 INCLUDING CLUSTER RANDOMISED DESIGNS IN META-ANALYSIS

3.2.1 Systematic reviews and meta-analysis

A systematic review generally aims to answer a research question or questions using existing literature via a systematic approach. Although no universal definition of a systematic review exists, common elements include a pre-specified protocol with strict inclusion criteria, a transparent literature search, quality assessment and some sort of evidence synthesis (92).

Meta-analysis (93) is one method of synthesising quantitative data in a systematic review. Meta-analysis is a set of statistical techniques for combining two or more reported summary statistics into a single estimate. Commonly the summary of interest is the intervention effect from RCTs. A systematic review of RCTs with meta-analysis is often cited to be the 'gold standard' of clinical research and the pinnacle of the hierarchy of evidence

(94). However, the approach is not suitable for every research question and is not without risk of bias (95).

Meta-analysis of RCT data can be considered as a weighted average of the treatment effects from individual trials (72), which ideally will be accompanied by a confidence interval (96). The standard method of doing this is the inverse variance method and the form of the assigned weights will vary depending as to whether the analyst has chosen a fixed or random approach.

A fixed effects meta-analysis relies on the assumption that the estimates from each study are estimating the same single underlying treatment effect (97). If T_i is the treatment effect estimated in the i^{th} study and SE_i is the standard error of that estimate then the weighted average T_f is calculated using *Equation 1*

Equation 1 (72)

$$\frac{\sum T_i / SE_i^2}{\sum 1 / SE_i^2}$$

A random effects meta-analysis uses the different assumption that the estimates from each study are estimating treatment effects that vary but follow the same underlying distribution. Here we will assume that the underlying distribution is normal with estimated mean T_r and estimated standard deviation σ . If T_i is the treatment effect estimated in the i^{th} study and SE_i is the standard error of that estimate, with between study variance σ then the weighted average T_r is calculated as

Equation 2 (97)

$$\frac{\sum T_i / SE_i^2 + \sum T_i / \sigma^2}{\sum 1 / SE_i^2 + \sum 1 / \sigma^2}$$

Statistical software such as R and Stata include meta-analysis packages, and the Cochrane Collaboration have their own bespoke systematic review software, RevMan which performs basic meta-analysis. These allow the user to input raw summary data (such as number of events, means, standard deviations) which are used to calculate the relevant treatment effect sizes and standard errors. Alternatively the user can input effect estimates and standard errors themselves; this is known as the generic inverse variance method.

3.2.2 Meta-analysis of CRTs

The inclusion of CRTs introduces additional challenges to a meta-analysis. Ignoring the clustering in a meta-analysis may lead to inappropriate weights and estimates that are over-precise, both of which could lead to incorrect conclusions.

When a CRT has been analysed with adjustment for clustering, and reported in sufficient detail, the resulting treatment effect estimate and the adjusted standard error associated with it should be extracted (adjusted).. The CRT can be included in meta-analysis alongside RCTs using the inverse-variance approach, above (72).

However, as discussed in Chapter 2, it is common for authors of CRTs to report effect estimates with standard errors and confidence intervals resulting from an analysis that has not been adjusted for clustering (unadjusted). The Cochrane handbook (72) describes a method to estimate the adjusted standard errors using an inflated standard error approach, using the design effect described in Chapter 2. Where an unadjusted standard error has been reported (or calculated by the systematic reviewer based on summary statistics), the inflated standard error is calculated by multiplying the unadjusted standard error by the square root of the design effect. The calculation of the design effect requires the ICC; where this is not reported the handbook recommends estimating using the ICC from similar trials. Where these are unavailable, ICCs can be estimated based on empirical studies or using a range of plausible ICCs (see section 2.2.1). Once inflated standard errors have been estimated, effect estimates can be combined using the generic inverse-variance method. This method is particularly useful when the systematic reviewer wishes to combine CRTs where clustering has not been taken into account alongside individually RCTs, and CRTs with correct adjustment for clustering. This method generally relies on estimated ICCs and assumes equal cluster sizes and therefore may be inaccurate.

The Cochrane Handbook also describes a second approach for including unadjusted estimates from CRTs in meta-analysis using an 'effective sample size' (98). This is calculated by dividing the sample size in both the intervention and control arms by the design effect, and rounding to the nearest whole number. For binary data, both the number of events and the group size should be divided by the design effect. Due to rounding and the limitations described in the paragraph above, this method is likely to lead to inaccurate effect estimates when the sample size is small so should be used cautiously. This method is particularly useful when the systematic reviewer wishes to enter raw summary data for all trials rather than effect sizes and standard errors.

Both methods above lead to an effective reduction in the weight assigned to the results from a CRT in a meta-analysis when compared to an analysis that ignores the clustering. Both methods only provide an estimate of the true weighting because they assume a fixed cluster size, and may be based on an estimate of the ICC; a properly adjusted estimate is preferable. Sensitivity analyses can be used to test the robustness of results based on these estimates are; e.g. using alternative ICCs.

3.2.3 Issues in the meta-analysis of CRTs

Although there is good guidance on how to include CRTs in meta-analysis, evidence from systematic reviews would suggest that these are under used.

A 2003 review (99) of meta-analyses including CRTs found that of 25 meta-analyses, only 3 (12%) attempted to account for clustering in their analysis while 6 (24%) reported the cluster randomised results separately and 15 (60%) included the CRTs as though they were individually randomised.

A 2016 review of Cochrane reviews (100) found that of 50 systematic reviews that included CRTs, only 28 (56%) mentioned CRTs specifically in the eligibility criteria of their protocol and only 8 (16%) reported methods of meta-analysis that took into account the cluster randomised designs.

While the methods described in this section address how to include parallel group CRTs in a meta-analysis they can't easily be applied to alternative cluster designs such as CRXO designs and stepped wedge trials.

3.3 ALTERNATIVE CLUSTER DESIGNS

3.3.1 Different designs

Alternative designs to the two arm parallel group CRT are the CRXO design (9), the stepped wedge trial (89) and CRTs with a baseline period (87); these types of alternative design need special treatment. Appropriate analysis of these designs incorporates both the between arm comparison (as in a parallel group trial) and the within arm comparison (comparing intervention and control periods in the same cluster) which may offer increased statistical power. When these trials have been analysed and reported adequately then effect sizes and standard errors can be utilised in meta-analysis using the generic inverse

method as described in 3.2.1. However, when analysis has not appropriately taken the design into account or when reporting is incomplete the application of methods described in 3.2.2 would focus on the between cluster comparison only and not take into account the within cluster comparison this resulting in a loss of precision.

3.3.2 The CRXO design

The CRXO design (9) has two or more time periods. In the case of two interventions and two arms, a cluster will receive either intervention or control first and then switch to the alternative treatment in the second period, but this can be extended to offer multiple interventions over multiple time periods (101). The CRXO trial can be a cross-sectional design where each period contains a different set of individuals, a cohort design where the same individuals are followed through all time periods, or a mixture of the two. There may be a washout phase between periods to allow for any carryover of intervention effects.

A CRXO design is only suitable for certain situations (101). It must be possible for a cluster to switch between intervention and control and back again without any long term carryover of effects; this would not work if the intervention were dependent on staff training for example, as after being part of an intervention period it would be impossible for staff to unlearn the new knowledge. A washout phase can be used between periods to allow for any short term carryover effects, for example to allow time for a pharmaceutical intervention to leave the body or for patients treated in the previous period to leave an ICU. It must also be quick and easy to switch between intervention and control phases, especially when there are more than two time periods, for example it would be difficult to run a CRXO trial if the intervention requires construction work, such as installing Perspex screens to minimise exposure to COVID-19.

The FLUID trial (102) is an example of the CRXO design. Hospitals were randomised to giving either saline or Ringer's lactate for fluid resuscitation. The hospital provided one product during an initial 12 week time period, and then switched to the alternative treatment for the second time period, with a 3 week washout period in between to allow changeover of stock. It is challenging to recruit individual patients to a trial of fluid resuscitation as it tends to be required very quickly and for critically ill patients.

Recruitment of entire hospitals (with a waiver of individual consent) and the use of routine health records allowed efficient and unbiased outcome assessment. Implementation of intervention (or control) simply required the switching of available product in all wards

within the hospital, so this was reasonably quick without long term carry-over or staff training.

Random effects models are common for the analysis of CRXO trials. These can take into account the multilevel data structure and should include 'cluster' as either a random or fixed, effect commonly with time period taken into account as a fixed effect (103). Where multiple measurements from the same participant are taken, e.g. in a cohort design, then a random effect for 'person' can be nested within cluster. Correlation structures can be specified; for example by assuming that correlation between participants from the same cluster will decay as clusters become further apart in time(104). Similarly, GEE methods can be adopted by specifying correlation matrixes with correlation within clusters and between time periods (105). A method that aggregates cluster level data can also be adopted – e.g. by calculating summary data from control and intervention periods relating to the same cluster, then either using paired methods or calculating differences and using weighted regression analyses (106). Note that the aggregate methods do not take into account temporal effects.

3.3.3 Including CRXO designs in meta-analysis

When it comes to incorporating CRXO designs in meta-analysis the literature is sparse. The Cochrane Handbook simply says 'The analysis of a cluster crossover trial should consider both the pairing of intervention periods within clusters and the similarity of individuals within clusters' and 'review authors are encouraged to seek statistical advice' (72).

Methods for CRTs described in section 3.2.2 can be applied to CRXO trials but they require either reported standard errors from an analysis that takes the CRXO design into account, or the application of a design effect. Design effects can be calculated for CRXO trials (107, 108); in addition to the ICC and cluster sizes they require an estimate of the correlation between two measurements from the same cluster at different time points and little guidance exists on how to estimate this.

3.4 METHODS APPLIED TO COCHRANE REVIEW OF CHLORHEXIDINE BATHING

3.4.1 Background

Lewis (8) is a Cochrane review which aims to assess the effectiveness of chlorhexidine bathing on hospital-acquired infections in people who are critically ill. This is a situation

where a CRXO trial is a valid design. The intervention is something that is quick and easy both to implement and to remove, it simply involves adding a chlorhexidine agent to the usual bathing process. The setting for critically ill patients is generally an ICU or critical care unit. These are settings where the turnover of patients is fast, and the outcome of infection tends to happen in a relatively short time, so carryover can be avoided by ensuring that the participants in each phase are different. Four trials in the review use a CRXO design, all of them cross-sectional. Note that if any of the studies had been CRXO cohort designs then it would have been important to allow for both within persons and within cluster correlation in the analysis.

Climo (109) is one trial in the review. Nine intensive care units were randomised to using either no-rinse chlorhexidine-impregnated washcloths or non-antimicrobial washcloths for an initial phase of six months after which they switched to using the alternate product.

I first became involved in this review as a peer reviewer. I noticed that the reviewers were using the 'effective sample size' method for all CRTs, utilising raw data on the number of events; this method was not incorrect but it was wasting statistical power by not fully utilising the CRXO results. For CRXO trials, the within cluster comparison was entirely ignored, which meant that the weight assigned to a large, well conducted robust trial would be smaller than it should be. Clustering would be expected to inflate the standard error, but the within cluster comparison would be expected to reduce it and lead to more precise estimates of treatment effect. By ignoring the correct standard errors provided by the trial authors small low quality trials were being given more weight that they ought to while high quality CRXO trials were penalised. I provided detailed peer reviewer comments that highlighted the available information in each trial, with suggestions on how best to utilise this. This resulted in me being invited to be a co-author to run these analyses myself.

3.4.2 Appropriate estimation of hospital acquired infection rates

Once I had joined the review teams I made a number of amendments

- (a) Extracting extra information on design, definition of outcomes, analysis methods and results from trial papers
- (b) Choice of 'rate differences' as an effect measure that allowed us to combine the maximum number of trials in meta-analysis
- (c) Methods to allow us to estimate adjusted effect sizes and standard errors for CRXO trials

- (d) New meta-analyses that incorporate CRXO trials efficiently
- (e) Sensitivity analyses to test assumptions
- (f) Update of review text and GRADE summary

The planned primary outcome measure for the review (110) was ‘hospital acquired infection’ summarised using odds ratios. This would require a binary ‘had at least one infection’ response for each person. On looking at the trial papers I noticed that many of the papers had reported infection outcomes as rates per patient day, which takes into account multiple infections and the time in the study. After discussion with the other review authors we decided to use ‘rate difference per 1000 patient days’ to summarise the primary outcome, and highlighted this protocol deviation. This would allow most studies to be reported on a similar scale, calculating rates using raw data or estimating rates from other available information. It also used the full information on rates (number of infections per unit of time) rather than collapsing this to a binary variable (whether or not a person has had at least one infection).

I looked at the reported data on ‘hospital acquired infections’ in each trial, and the method of analysis utilised. In each I case tried to extract an estimate of the rate difference and its standard error. Wherever possible I used standard errors that properly take into account both clustering and the correlation between repeated measurements from the same cluster. I considered a CRXO trial to have taken into account the CRXO design if it made reference to any of the methods described in section 3.3.2 with explicit mention of cluster adjustment. This information is summarised in Table 3-1.

Two CRXO trials (out of four) (111, 112) had reported a rate difference with 95% confidence interval from an analysis that considered the CRXO design and. In these cases the standard error could be estimated from the confidence interval (assuming a confidence interval of 1.96 standard errors either side of the effect estimate).

Three individually RCTs (113-115) reported the number of infections and number of patient days. This allowed rate differences and standard errors (SE) to be calculated using *Equation 3* and *Equation 4* below where E_i and E_c are the number of events (infections) in the treatment and control group and T_i and T_c are the number of patient days (72). These formulae assume that the events follow a Poisson distribution (with equal mean and variance) and do not allow for over-dispersion. *Equation 3* (72)

Equation 4 (72)

Two trials (109, 116) (one CRXO trial and one parallel group trial) reported rates but without a standard error or confidence interval; they reported only p-values. I looked up z-values corresponding to the p-values and used these to estimate the standard errors by dividing the calculated rate difference by the estimated z-value. This is only a crude estimate of the true standard error because it assumes that the p-value comes from a simple z-test which is unlikely to be true; it is likely it comes from a statistical model with adjustment for other factors. Note that given the small number of clusters, use of the t-distribution would have been more appropriate here(76).

The final CRXO trial (117) reported a rate ratio with 95% CI as well as raw data on infections and time. Methods to estimate adjusted standard errors for a rate difference are described in section 3.4.3 below.

3.4.3 Design effects in CRXO trials

For a CRT, the design effect (DE) is based on the degree of clustering and the size of cluster. Empirical research suggests that the ICC is reasonably consistent across trials from similar settings and with similar cluster sizes (118). This would suggest that the design effect ought to be consistent from one format of an outcome to another within the same trial (e.g. when converting from a rate ratio to a rate difference) – due to equal cluster sizes and a similar ICC –. In the absence of additional information, I made an assumption that the same would be true for CRXO trials, however this may not be true as the design effect for CRXO trials will also depend on the within cluster correlation (108) so further research is required to test this assumption.

A formulae for design effect for a simple 2 period CRXO trial is given by Hooper et al. (107). This is obtained by multiplying the design effect relating to cluster randomisation by a

design effect for repeated measures from the same cluster $(1-r)/2$ (where r is the correlation between 2 measurements from the same cluster at different time points) (Equation 5). Calculation of the design effect requires estimates of both the ICC and r which were not reported for any of the trials in the Chlorohexadine review.

Equation 5 (107)

Design effect for a CRXO trial where m = cluster size, r = correlation between repeated samples from the same cluster and ρ = ICC

—

Milestone(117) was a CRXO trial with 10 clusters and 4947 participants, which was analysed using Poisson regression with adjustment for cluster and time to take into account the CRXO design. This trial reported both crude unadjusted rate ratios and CIs as well as adjusted ones. They reported 28 central-line associated bloodstream infections over 9333 patient days in the control arm and 13 infections over 7975 patient days in the control arm.

Reported data described the rate of central-line associated bloodstream infections (per 1000 patient days) 3 vs 1.63, crude incidence rate ratio 0.54 (95% CI 0.26 to 1.08) adjusted incidence rate ratio 0.52 (95% CI 0.25 to 1.08).

I calculated natural logs of the reported rate ratio and 95% confidence interval and used these to calculate the unadjusted and adjusted standard error of the

Unadjusted $\ln(\text{rate ratio})$; $\ln 0.54 (0.26 \text{ to } 1.08) = -0.616 (-1.337 \text{ to } 0.077)$ Width of confidence interval = $0.077 - (-1.337) = 1.414$. Standard error = $1.414/1.96 = 0.721$.

Adjusted $\ln(\text{rateratio})$; $\ln 0.52 (0.25 \text{ to } 1.08) = -0.654 (-1.386 \text{ to } 0.077)$ Width of confidence interval = $0.077 - (-1.386) = 1.463$. Standard error = $1.463/1.96 = 0.746$.

Once I had both the adjusted and unadjusted SE of the \ln rate ratio I divided the adjusted value by the unadjusted value ($0.746/0.721$), giving the square root of the VIF to be 1.03 (Equation 6) (and design effect of $1.03^2 = 1.06$). This value seemed plausible, while the clustering would be expected to inflate the standard error, the within cluster crossover would be expected to bring it back closer to 1. Note that the trial authors estimated a design effect of 1.2 when calculating sample size so this is the same order of magnitude.

Equation 6 (72)

Utilising the raw data and the formulae for rate difference and SE (*Equation 3* and *Equation 4*) I calculated the rate difference = $(28/9333)*1000 - (13/7975)*1000 = 1.370$, and SE = $\text{sqrt}((28/9333^2) + (13/7975^2)) = 0.000725$. This is the standard error for time per patient day, so this was multiplied by 1000 to convert to units of 'per 1000 patient day' to give 0.725. Multiplying the unadjusted SE by the square root of the VIF $0.725*1.03$ gave an estimate of the adjusted standard error as 0.745 that could be utilised in generic inverse meta-analysis.

Table 3-1 Methods for estimating rate difference (and 95% confidence interval) for the Chlorhexidine bathing systematic review

Study	Study design	Analysis reported by study author	Data reported in trial report	Data manipulation
Bleasdale 2007 (111)	Cluster-randomised cross-over trial with 2 clusters and 836 participants	CRXO design taken into account with multivariate models that included a fixed term for geographical unit. Unclear what multivariate model was used.	Rate (per 1000 patient days) 10.4 vs 4.1, 95% CI for rate difference 1.2 to 11	SE for rate difference calculated from CI
Boonyasiri 2016 (113)	Parallel group trial of 481 participants	Individual incidence	28 infections during 3284 patient days vs 29 infections during 2759 patient days (adding up	Rate difference and associated SE calculated from summary data on infections and patient days

			infections and using mean ICU stay to calculate patient days)	
Camus 2005(114)	2x2 factorial trial with 256 participants in two relevant arms	Number of infections and patient days	87 infections during 1961 patient days vs 87 infections during 1991 patient days	Rate difference and associated SE calculated from summary data on infections and patient days
Climo 2013 (109)	Cluster-randomised cross-over trial with 9 clusters and 7727 participants	CRXO design taken into account with GEE (according to SAP)	Rate (per 1000 patient days) 6.60 vs 4.78, P = 0.007	P value used to calculate Z-value, and rate difference/ Z gives estimate of SE
Milestone 2013 (117)	Cluster-randomised cross-over trial with 10 clusters and 4947 participants	CRXO design taken into account with Poisson regression adjusted for cluster and time	Rate (per 1000 patient days) 3 vs 1.63, incidence rate ratio 0.52 (95% CI 0.26 to 1.08), adjusted CI 0.25 to 1.08	Unadjusted and adjusted CI used to calculate SE for log rate ratio: from these design effect = 1.03^2 applied to inflate SE of rate difference
Noto 2015 (112)	Cluster-randomised cross-over trial with 5 clusters and 9340 participants	CRXO design taken into account in supplementary materials which present group level analysis of clusters	Rate (per 1000 patient days) 3.35 vs 3.31, 95% CI for rate difference (-1.19 to 1.11)	SE for rate difference calculated from CI

Pallotto 2018 (116)	Parallel group trial of 449 participants	Number of infections per patient days	Rate (per 1000 patient days) 40.9 vs 23.2, P = 0.034	P value used to calculate Z-value, and rate difference/ Z gives estimate of SE
Swan 2016 (115)	Parallel group trial of 350 participants	Hazard ratios and risk difference	35 infections during 2416 patient days vs 18 infections during 2332 patient days (supplementary digital content)	Rate difference and associated SE calculated from summary data on infections and patient days

GEE: general estimating equation; ICU: intensive care unit; SAP: statistical analysis plan; SE: standard error.

Similar methods were applied to the secondary outcome of mortality, and these are detailed in the review appendix (8,Appendix 9).

3.4.4 Combining estimates of hospital infection rates

Figure 3-1 shows a forest plot for the meta-analysis of hospital acquired infections from the 8 trials in the review. I chose to present the parallel studies and cluster-randomised crossover trials as separate subgroups. Several authors warn that different trial designs can lead to heterogeneity in effect size (99, 119, 120) and highlight the possibility of an interaction between the unit of randomisation and the treatment effect. On average the parallel group trials tended to show larger effect sizes (rate difference 4.00(95% CI -3.14 to 11.4)) than the CRXO trials (rate difference 1.41(95% CI 0.00 to 2.83)) , but the parallel group trials were also on average much smaller and therefore more imprecise; the test for subgroup differences did not confirm a difference. Combining all 8 trials suggested a reduction in rate of 1.70 (95% CI 0.12 to 3.29) infections per 1000 patient days when using Chlorohexidine compared to soap and water.

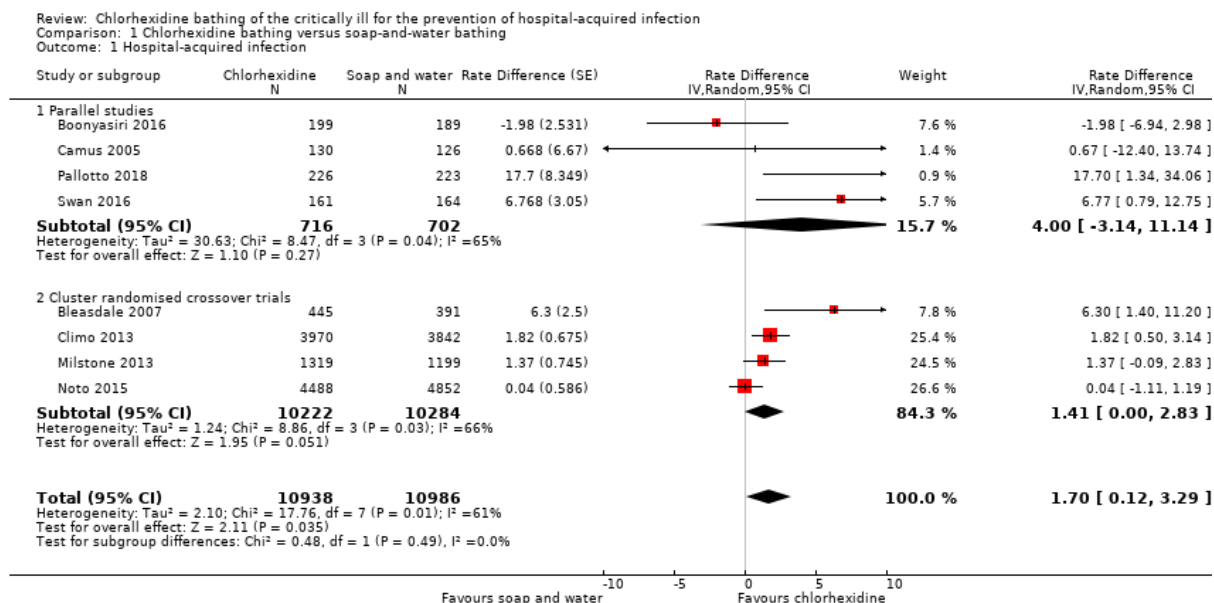


Figure 3-1 Forest plot and meta-analysis comparing chlorhexidine bathing to control

3.4.5 Sensitivity analyses

Bleasdale (111) has only 2 clusters. There was no minimum sample size or number of clusters specified in the protocol for the systematic review so at the time we made the decision not to exclude this trial but it is clear that a trial with only 2 clusters has very poor external validity and a high risk of type 1 error (false statistical significance) (76). In addition, Bleasdale 2007 and Swan 2016 have incongruously large risk differences of 6.30 and 6.77. Sensitivity analyses reported in *Table 3-2* show that removing either of these studies would reduce the overall risk difference sufficiently to cause conclusions to change and we downgraded the level of evidence using GRADE to reflect this

The parameters for Millstone 2013 were based on an estimated design effect; I explored the effect of using a more conservative design effect based on standard formulae for a parallel CRT; after doing this the confidence interval was no longer entirely above the null line, meaning that we cannot be certain whether or not Chlorhexidine bathing is effective.

These sensitivity analyses suggested that the results were not robust and should be interpreted cautiously. We used the GRADE approach, as planned, to downgrade the certainty of the evidence by one level based on the sensitivity analysis, with another downgrade based on concerns about risk of bias. This meant that the review concluded

that we were uncertain about whether or not chlorohexidine bathing reduced the risk of infection.

Table 3-2 Results of sensitivity analyses for the Chlorohexidine systematic review

Analysis	Risk difference (95% CI)
Including all studies	1.70(0.12 to 3.29)
Removing Bleasdale 2007	1.26(-0.21 to 2.72)
Removing Swan 2016	1.35(-0.15 to 2.85)
Removing Bleasdale 2007 and Swan 2016	0.96(-0.35 to 2.27)
Using a conservative design effect for Millstone 2013	1.97(-0.06 to 4.00)

3.5 RECOMMENDATIONS FOR COMBINING CRXO TRIALS IN META-ANALYSIS

- (1) An appropriately analysed CRXO trial will have had an analysis that adjusts for clustering and takes into account both within and between cluster comparisons. Wherever possible researchers should try to extract and utilise reported effect sizes and adjusted standard errors from trial reports.
- (2) Effect sizes and adjusted standard errors from a CRXO trial can be combined alongside individually randomised and parallel CRTs using a generic inverse meta-analysis. Be aware that using methods that ignore the robust within-cluster comparisons (e.g. by following effective sample size methods suitable for parallel group CRTs) wastes information which is unethical as it ignores the research contribution of the individuals involved. Downplaying evidence from this robust trial design in the review could potentially lead to bias.
- (3) When intending to combine CRXO trials with other trial designs in the same meta-analysis it is advisable to display them as separate subgroups (e.g. individual RCTS, CRTs, CRXO) and consider any heterogeneity due to design carefully.

- (4) It may be necessary to manipulate outcome measures to all adhere to a common format (e.g. where trials have a mixture of RRs and ORs). When considering which summary measure to use, it may be necessary to prioritise the summary measures used in reports of CRXOs to avoid estimation of standard errors that may be inaccurate. Note that if this necessitates a deviation from the systematic review protocol, this should be documented clearly.
- (5) There are situations where adjustment for the CRXO design will need to be applied using a design effect e.g. when manipulating outcome measures from CRXO trials from one format to another or where appropriately adjusted standard errors have not been reported. Design effects can be estimated from (a) very similar trials with the same outcome or (b) from different outcomes within the same trial. This should be done cautiously. Further research is required to look at how best to estimate design effects for CRXO trials with incomplete reporting.
- (6) Where standard errors have been estimated, it is important to use sensitivity analyses to test whether results are robust, for example by using a range of alternative design effects, excluding the CRXO trial from the meta-analysis or by using formulae for CRTs (down-weighting the CRXO trial by ignoring within cluster comparison).

3.6 REFLECTIONS

I was delighted to get the opportunity to work on such a methodologically challenging systematic review. When I provided comments as a peer reviewer I had no idea how unusual it was to find cluster crossover trials in a systematic review, or how limited the guidance is on their inclusion in a meta-analysis. I feel that I added something really useful to this review, partly in recognising the appropriateness of the CRXO trial in this context, and also in ensuring that these trials received the appropriate weight in meta-analysis.

This work has highlighted for me the potential uses for CRXO trials and on several occasions I have advocated their use in statistical consultancy. I intend to apply for some funding via either a fellowship or an MRC methodology grant to explore how to estimate a design effect specific to CRXO trials, and further develop guidance on suitable methods for

incorporating CRXO trials in meta-analysis. This would likely include a consensus study with meta-analysts and some simulation work.

3.7 METRICS

Paper 3 is published in the Cochrane database of systematic reviews. It has 14 citations and parts of it have been translated into 10 different languages. This Cochrane Review has an Altmetrics Attention score of 55 which puts it in the top 5% of research outputs reviewed by Almetrics. It also appears in a Cochrane 'Clinical Answers' article (121) and a Cochrane Special Collection of Coronavirus infection control and prevention measures (122).

CHAPTER 4 INCORPORATING CRTs INTO A META-ANALYSIS OF COMPLEX BEHAVIOUR CHANGE INTERVENTIONS

4.1 CHAPTER OVERVIEW

In chapter 3 I covered how to incorporate CRTs into meta-analysis, including non-standard cluster randomised designs. In this chapter I discuss how to incorporate CRTs into systematic reviews with additional layers of complexity; mixed interventions, mixed layers of clustering and mixed outcome measures, using the example of a systematic review of a complex behaviour change intervention, the SOCIAL review. Paper 4 is the protocol of the SOCIAL review, Paper 5 is the published report to the funding body (NIHR Health Services and Delivery Research Panel) and Paper 6 is a journal article summarising results. Paper 7 is a comparison of different evidence synthesis methods applied to this review. While some of the problems discussed aren't exclusively limited to CRTs, they will demonstrate that dealing with clustering is one of a package of items to consider when attempting to meta-analyse pragmatic trials of complex interventions.

4.2 SYSTEMATIC REVIEWS OF COMPLEX INTERVENTIONS

4.2.1 Complex behaviour change interventions

A complex intervention is defined to consist of multiple components or be dependent on multiple factors (123). In this chapter, I will use the example of the SOCIAL systematic review (14, 15, 124). The trials within the SOCIAL review had both of these features; interventions which generally consist of several active components and whose effectiveness is very likely to vary by setting and/or context.

4.2.2 Challenges for systematic reviews of complex interventions

Systematic reviews of complex interventions face number of challenges. Trials of complex interventions tend to be very heterogeneous in terms of the intervention, the way it is implemented, the setting and the trial design (125, 126); this leads to statistical complexity and heterogeneity and difficulties in interpretation.

A complex intervention could be a mixture of beneficial, ineffective or even harmful ingredients. It is important to decide whether the research questions will address the overall effectiveness of a package of intervention components or try to isolate the effect of individual components (127, 128). When considering a set of individual components it may also be important to know they interact with each other; i.e. whether or not the effectiveness of a component varies according to the other interventions that it is partnered with.

A review of complex interventions is often about more than whether the intervention works, but also the mechanism of how it works (129) and under what circumstances it works best.

4.2.3 Methods suitable for systematic reviews of complex behaviour change interventions

A number of approaches to synthesise quantitative data do not use meta-analysis.

Graphical methods include the albatross plot (130), the harvest plot (131) the bubble plot (132). A qualitative comparative approach (133) looks for effective interventions and then uses set-theory to look at which features (intervention components, length of time, mode of delivery) they have in common. These methods do not require effect estimates or their standard errors so they can be used when this information is unavailable or the outcomes have not been reported in a consistent format (132). They may also be useful to display results when the studies are considered far too heterogeneous to combine in meta-analysis. However, while these methods provide a useful summary, they may increase the risk of bias and are unlikely to allow firm conclusions.

Several authors (126, 134, 135) cite the need to combine both quantitative and qualitative evidence when considering complex interventions. Mixed methods approaches include thematic (136), realist (137) or framework (138) synthesis methods. Incorporating data on from interviews and questionnaires may reveal information about the mechanism of change, reasons for variation or barriers to implementation, for example. However, the methods are less developed than those for a purely quantitative review, and approaching questions about effectiveness may be difficult to approach objectively.

Network meta-analysis (139) is an extension of the standard pairwise meta-analysis described in Chapter 3 where more than two interventions can be evaluated. When analysing a set of competing interventions, network meta-analysis may lead to more statistical power than multiple separate pairwise analyses because it utilises all available

information. It also ranks the interventions in order of effectiveness, even when there is no direct trial evidence for some pairs of interventions.

Components based meta-analysis (125) is a meta-analytic approach that aims for the effect of each component within a complex intervention to be evaluated, along with any interactions between them. This method is useful when interventions to be split into a number of 'clinically meaningful units' and the aim is to understand the usefulness of each.

Subgroup analyses (72) involve splitting data for subsets of studies (e.g. different settings, different types of intervention). Subgroups can be presented separately on forest plots to investigate heterogeneous results, or to answer questions about particular groups.

Researchers conducting meta-analysis generally have a large number of different subgroups available so to avoid multiplicity issues and data dredging it is important to specify these in advance as far as possible.

Meta-regression (140) allows a set of trial level covariates to be included in meta-analysis. Taking into account factors relating to setting, population, trial design and intervention is highly desirable in a meta-analysis of complex interventions as these features are known to increase heterogeneity. Furthermore, knowing how the effect of the intervention varies for different types of patient or for different variants of the same intervention is often a key research aim. However, meta-regression is often underpowered and it relies on observational data and so is at high risk from confounding. Researchers conducting meta-regression should specify covariates in advance.

4.3 CHALLENGES FACED IN THE SOCIAL REVIEW

4.3.1 Description of the SOCIAL systematic review

The SOCIAL review aims to answer questions about whether or not social norms interventions are effective at changing the clinical behaviour of healthcare professionals. The behaviour change taxonomy (141) is a classification system for behaviour change techniques (BCTs) and identifies 93 individual techniques that could be part of a behaviour change intervention. We recognised that 5 of these BCTs could be considered to have a social norms element whereby they are designed to change behaviour by utilising "implicit or explicit behavioural rules that one uses to determine the appropriate and/or typical expectations, beliefs, attitudes and behaviours of a social reference person or group" (15,p1). The 5 social norm BCTs were social comparison, information about other's approval, credible source, social rewards and social incentive. However, when looking for

trials of interventions containing a social norms BCT, it was rare to find that the social norms BCT was the only BCT in the intervention; most interventions were a bundle of different BCTs all aiming to actively change behaviour. For example, a trial designed to improve the quality of care for patients with diabetes (142) had an intervention which included both a social norm element where clinicians were provided with graphical displays that allowed them to compare their own results with neighbouring practices (social comparison) and pop-up patient information to use during consultations (prompts and cues).

Table 4-1 *Challenges faced in the SOCIAL systematic review* outlines the main challenges faced in the synthesis of data within the systematic review, the method that I chose to address these and a signpost to where this is described.

Table 4-1 Challenges faced in the SOCIAL systematic review

Challenge	Methods chosen	Section
Interventions were a mixture of different bundles of BCTs	Group be commonly occurring packages. Meta-regression. Network meta-analysis.	4.4.1
Control arms were a mixture of different bundles of BCTs	Subtracting of BCT interventions	4.4.2
Inclusion of multi-arm studies	Careful consideration. Treat each 'comparison' as a separate unit.	4.4.2
Varied level of clustering	Appropriate adjustment for unit of randomisation. Sensitivity analyses using alternate approaches.	4.4.3
Mixed outcome measures	Use of standardised mean differences. Sensitivity analyses using alternate approaches.	4.4.4

4.4 METHODS CHOSEN FOR THE SOCIAL SYSTEMATIC REVIEW

4.4.1 Meta-analysis methods in the SOCIAL systematic review

In the SOCIAL systematic review we had the following research questions;

1. “What is the effect of interventions containing social norms BCTs on (a) the clinical behaviour of healthcare workers, and (b) resulting patient health outcomes?”
2. Which contexts, modes of delivery and behaviour change techniques are associated with the effectiveness of social norms interventions on healthcare worker clinical behaviour change?” (15,p3)

To answer the first question, the ideal plan would be to combine in one meta-analysis all studies that compare ‘any social norm’ intervention to a control group. The social norms interventions would be split into subgroups by the type of social norm BCT (e.g. credible source) to look at the consistency of results and see which of the social norm BCTs performed best, if any. The reality of behaviour change interventions is that they are almost always part of a complex intervention which contains multiple BCTs. In our protocol our inclusion criteria was any RCT with at least one social norm BCT in its intervention arm (124). This included trials with social norm BCTs in both arms as these would help to answer our second question.

Colleagues planned to code all BCTs in the intervention and control arms of all trials in the review. When planning the analysis for this project I had two main approaches that I thought were suitable for dealing with the data (a) grouping together sets of BCTs that occurred commonly together and (b) a components based approach, looking at the effect of each BCT separately.

The components based approach seemed useful initially, but as I read about it and looked at other studies, I realised that our search strategy and inclusion criteria meant that it was not really suitable. An example of a components based meta-analysis is a review comparing different psychological preparation interventions for adults undergoing surgery. This review include any psychological intervention and these were thought to consist of one or more of 5 key components; reviewers were able to draw a network linking all of the pairwise

comparisons between any individual component or combination of components (143, Fig 1). My concern was that we were including only trials that contained social norm BCTs in at least one intervention. If we tried to draw a network treating both the social norm BCTs and other concomitant BCTs as 'meaningful components' then we'd be missing all the comparisons between the non-social norm components, which could lead to incorrect conclusions. I argued for a grouping approach instead where we looked for commonly occurring packages of interventions.

I planned a pairwise approach initially to answer our first research question; this meant that we only used trials that compared a social norm to a control (no intervention, standard practice or a non-social norm intervention). Subgroup analysis was used so that we could answer an overall question about social norm interventions in general, while also looking at which of our social norm packages appeared most effective.

I then extended this to a network meta-analysis. This allowed us to add in trials that compared two different social norm interventions. It also allowed us to rank the 'social norm packages' in order of effectiveness.

In addition I also decided to use meta-regression to answer question 2, with the caveat that we had to approach the results very cautiously.

4.4.2 Comparisons in the SOCIAL review

A challenging aspect of the SOCIAL review was trying to establish what exactly each trial or comparison within a trial was trying to test. The control arm would quite often contain either a subset of the BCTs in the intervention arm or alternative BCTs. I identified that simply grouping together all similar 'social norm interventions' and comparing them to any 'control' was not going to work. At the data extraction stage I asked the team to try to identify which of the following definitions best fit each trial

- (a) Social norm intervention v control
- (b) Social norm intervention + X v X
- (c) Social norm A v Social norm B

However, this proved difficult. Partly because the team were struggling to identify multifaceted concomitant interventions (X), and partly because of the large number of multi-arm studies, including factorial trials.

In order to work out exactly what was being tested in each trial we adopted a 'subtraction' approach, subtracting the BCTs in the control arm from those in the intervention arm to see what the trial was testing (

Table 4-2 *Examples of trials in the SOCIAL systematic review, and how we established what the trial was testing using subtraction*). This assumes there is no interaction between BCTs (i.e. that type (a) and type (b) trials are both estimating the same treatment effect); perhaps this is a strong assumption. When looking for ‘social norm packages’ it was the set of BCTs left after the subtraction that we used in order to define the social norm package that the trial was aiming to test.

In addition, I suggested that our database needed to have one row per ‘useful trial comparison’ so we could utilise all available information. This meant that some trials had multiple rows. To avoid any double counting, adjustments were made when necessary (e.g. halving numbers in the control group if the same control group was used in two different comparisons) (72). Note that this method was chosen (rather than combining groups) as it allowed two different interventions to belong to two subgroups in meta-analysis however it ignores correlation between two results from the same study.

Table 4-2 Examples of trials in the SOCIAL systematic review, and how we established what the trial was testing using subtraction

Trial	BCTs in Intervention arm (A)	BCTs in control arm (B)	Difference (A-B)	Description of package being tested
Boet 2018 (144)	4.1 instruction on how to perform the behaviour 6.2 social comparison	4.1 instruction on how to perform the behaviour	6.2 social comparison	Social comparison alone
Lakshiminarayan (145)	1.3 Goal setting 2.2 Feedback on behaviour 9.1 Credible source	No identified BCTs	1.3 Goal setting 2.2 Feedback on behaviour 9.1 Credible source	Credible source with other BCTs

Factorial trials had to be considered separately because wherever possible I wanted to use the estimated effect of the social norm intervention, rather than considering each arm of the factorial trial separately. Meeker 2016 (146) is a 2x2x2 factorial trial in the SOCIAL systematic review, which is testing the effect of 3 behavioural interventions simultaneously. One of the interventions is an example of social comparison, with emails sent to clinicians comparing their antibiotic prescribing rates with others. The social comparison intervention is present in 3 arms of the trial but rather than considering each of the arms and the BCTs within them separately, I utilised the effect size and confidence interval relating to the effect of the social comparison intervention that was presented in the trial report. This had been obtained via a covariate in a multilevel linear model. In the analysis, I treated this as though it were the results from a single comparison comparing social comparison to an inactive control.

4.4.3 Levels of clustering

As described in Chapter 2, CRTs can be combined with individually RCTs using standard errors that have been adjusted for clustering. Other cluster randomised designs, including stepped wedge CRTs (89) can also be included in this way. This method was used throughout the SOCIAL review which included individually RCTs, CRTs and stepped wedge trials. Where the outcomes of interest were not already adjusted I adjusted them myself using an estimated ICC as described in Chapter 3.

When trying to identify CRTs in the SOCIAL review, discussions between the team revealed that this was by no means clear. A common definition of a CRT is ‘randomisation of groups (clusters) of individuals to control or intervention conditions’ (2). In health research, the ‘individuals’ are usually patients or members of the public; a common design would be randomising groups of patients served by a particular general Practitioner (GP) or in a particular hospital ward.

The target of inference in the SOCIAL review was health care professionals rather than patients (147). A trial looking at GP behaviour may randomise GPs, but then collect data on patients treated by the GP to assess whether or not the GP has changed his or her behaviour in the way that they have been treating them. This trial could be analysed at the level of the GP (e.g. taking in average over all patients, no clustering adjustment required) or at the level of patient (adjustment for clustering by GP). When analysed at the GP level, is a trial like this a CRT? Some of our team felt yes because groups of patients were being randomised and some felt no because there was individual randomisation of members of the population of interest. Regardless of nomenclature, the patients here are still subjects of the research and are clustered at a group level and therefore all the special considerations of ethics, bias, analysis and reporting relating to CRTs apply.

Additional complexity came from the fact that many trials randomised groups of healthcare professionals (e.g. GPs in a surgery, nurses on a ward). In these trials the unit of analysis could be patient, healthcare professional or group of healthcare professionals. For each trial I tried to identify the unit of randomisation and unit of analysis. Table 4-3 shows the unit of randomisation and analysis for 16 of the trials in the review that used Credible Source interventions.

Table 4-3 *Units of randomisation and analysis for trial that included credible source in their intervention*

	Number of studies
Unit of randomisation	
Patient	0
Health care professional	2
Site (ward, hospital, surgery etc.)	14
Unit of analysis	
Patient	8
Health care professional	4
Site (ward, hospital, surgery etc.)	4

These mixed levels of randomisation and analysis made it quite difficult to interpret combined results of these trials because they did not apply to a consistent population. Their differing designs were also likely to add to the heterogeneity. After consideration we decided, on my recommendation to use adjusted standard errors. Other authors (148) have used an alternative approach, weighting the trials in the meta-analysis by the number of healthcare professionals in the study. This approach may be useful when a trial has not reported standard errors from an appropriate model that adjusts for clustering however this approach has a number of issues. Weighting by the number of healthcare professionals ignores precision/variability and does not taking into account the level of randomisation. In paper 7 (17) I discussed this issue in further detail and compared the utility of the two approaches by applying them both to the same set of data from the SOCIAL review.

4.4.4 Standardised mean differences

The SOCIAL systematic review looked at interventions to change health professional behaviour; the target behaviours were very diverse and include handwashing, test ordering, antibiotic prescribing etc.; and the aim of the intervention could be to increase or decrease the number of times the behaviour is performed. The outcome measurement in the trial could be something binary (e.g. whether the behaviour was performed or not) or scale (the number of times the behaviour was performed, the proportion of times the behaviour was performed). Both the type of behaviours and the format of reporting varied greatly across trials. As we wanted to answer a broad review question about behaviour in general, I suggested the use of standardised mean differences (149), and this is the approach we took. A comparison with other methods (17) suggests that conclusions were reasonably robust when alternative methods were applied to the same set of data.

4.5 REFLECTIONS

I had worked on a number of systematic reviews before I started working on the SOCIAL project, and peer reviewed many more as a Statistical Editor for Cochrane. The SOCIAL systematic review was by far the most challenging. One of the members of the steering committee once said to me something along the lines of ‘Wow, you’ve really got it all in this review – a mixture of complex interventions, a mixture of outcome measures, a mixture of trial designs and lots of other heterogeneity too’. I worried at times about whether or not we should have been attempting meta-analysis in this review. It certainly seems unlikely that the assumptions of a common underlying treatment effect that is expected for a fixed effects meta-analysis would have been met. However, we were clear from the start that in this review we were using meta-analysis as a weighted average of the observed data rather than trying to make inferences about the future; and I made sure that our conclusions were cautious and hypothesis generating.

I think that a ‘one size fits all approach’ for systematic reviews can be unhelpful; while the level of heterogeneity seen here would not be acceptable in a systematic review trying to estimate the precise treatment effect for a drug, it seems better in this exploratory work to attempt to combine the data rather than conduct multiple separate analyses. I feel that our conclusions were robust; on average, in the trials we found, there was a small beneficial effect of social norms in general, and of the social norm interventions tested, those that included credible source tended to have the largest treatment effect. Reassuringly, when I

attempted alternative methods of data synthesis, the overall conclusions on the effectiveness of social norms and the specific conclusion for credible source remained robust.

I am disappointed that I did not have the opportunity to use components based meta-analysis in this project; I think this technique really has merit when trying to evaluate complex interventions. I am pleased that I undertook training on network meta-analysis through the project and apply this for the first time. This knowledge has already been invaluable as collaborator on a suite of Cochrane Reviews and an overview looking at measures to prevent pressure ulcers (150).

This review chose a 'fixed effects meta-analysis' to be primary and we have had a lot of questions about this given that it is such a heterogeneous review. When writing the protocol, this point caused huge debate amongst collaborators and colleagues and the decision was not taken lightly. A commonly held point of view is that a random effects meta-analysis is a more conservative approach and that it accounts for heterogeneity by allowing the estimated treatment effect to vary between trials according to a common variance (exchangeability) (151). However there are contra-arguments against using random-effects with suggestions that the underlying assumption of exchangeability may be implausible in many cases (152) and that there are situations where a random effects meta-analysis is actually less conservative (overly narrow confidence intervals) (153). The fixed effects is known to give more weight to the largest/most precise trials (154) and we'd expect these to be less prone to bias. In our case, we expected systematic differences between trials and therefore were using meta-analysis as a statistical summary of the available evidence (155) rather than expecting to fully parameterise the variability and make precise inference. We were also careful to report both fixed and random effects throughout to show whether or not results were robust to our decision.

Overall, I'm really proud of my work on the SOCIAL review. I was part of it from the beginning, my first success as co-applicant on an NIHR grant, and I had a large role in making sure that this project was delivered on time and to a high quality.

4.6 METRICS

Paper 4 is published in the journal 'Systematic Reviews' which publishes high quality systematic review protocols and reviews related to health and has a 5 year impact factor of 5.08 . Paper 4 has 2 citations and over 9500 accesses. Paper 5 is published in 'Implementation Science; it has 1 citation and an Altmetric Attention score of 18. Implementation science publishes research about methods relevant to healthcare in clinical, organisational, or policy contexts; it has a 5 year impact factor of 8.71. Paper 6 is published as a report by the project funders, the NIHR Health Services and Delivery Research. Paper 7 is a pre-print published by Authorea with 59 views.

A summary of the recommendations from the SOCIAL systematic review have been published by the Audit and Feedback Meta-Lab as part of their recommendations on designing audit & feedback interventions. <http://www.ohri.ca/auditfeedback/resources-a-f-recommendations/>

In 2021 I was co-applicant on an application for a Wellcome Trust Institutional Translational Partnership Award (TPA) Projects for Translation (P4T) about the use of a credible source feedback intervention to change the behaviour of GP practice staff to reduce the ordering of blood tests for fatigue. Although this was unsuccessful we plan to submit this to a suitable funder in future.

REFERENCES

1. Rhodes S, Wilkinson J, Pearce N, Mueller W, Cherrie M, Stocking K, et al. Occupational differences in SARS-CoV-2 infection: analysis of the UK ONS COVID-19 infection survey. *Journal of Epidemiology and Community Health*. 2022;jech-2022-219101.
2. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ*. 2017;358:j3064.
3. Patchwood E, Rothwell K, Rhodes S, Batistatou E, Woodward-Nutt K, Lau Y-S, et al. Organising Support for Carers of Stroke Survivors (OSCARSS): study protocol for a cluster randomised controlled trial, including health economic analysis. *Trials*. 2019;20(1):19.
4. Patchwood E, Woodward-Nutt K, Rhodes SA, Batistatou E, Camacho E, Knowles S, et al. Organising Support for Carers of Stroke Survivors (OSCARSS): a cluster randomised controlled trial with economic evaluation. *BMJ Open*. 2021;11(1):e038777.
5. Darley S, Knowles S, Woodward-Nutt K, Mitchell C, Grande G, Ewing G, et al. Challenges implementing a carer support intervention within a national stroke organisation: findings from the process evaluation of the OSCARSS trial. *BMJ Open*. 2021;11(1):e038129.
6. Sim J, Dawson A. Informed Consent and Cluster-Randomized Trials. *American Journal of Public Health*. 2012;102(3):480-5.
7. Farrin A, Russell I, Torgerson D, Underwood M. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clinical trials (London, England)*. 2005;2(2):119-24.
8. Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. *Cochrane Database of Systematic Reviews*. 2019(8).
9. Arnup SJ, McKenzie JE, Hemming K, Pilcher D, Forbes AB. Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. *Trials*. 2017;18(1):381.
10. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ : British Medical Journal*. 2012;345:e5661.
11. Higgins JPT, Cochrane C. *Cochrane handbook for systematic reviews of interventions* 2019.
12. Sanchez-Meca J, Marín-Martínez F. Weighting by Inverse Variance or by Sample Size in Meta-Analysis: A Simulation Study. *Educational and Psychological Measurement*. 1998;58(2):211-20.
13. Cotterill S, Powell R, Rhodes S, Brown B, Roberts J, Tang MY, et al. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. 2019;8:176.
14. Cotterill S, Tang MY, Powell R, Howarth E, McGowan L, Roberts J, et al. Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis. 2020;8:41.
15. Tang MY, Rhodes S, Powell R, McGowan L, Howarth E, Brown B, et al. How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis. *Implementation Science*. 2021;16(1):8.
16. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Stat Med*. 2002;21(19):2971-80.

17. Rhodes S, Dias S, Wilkinson J, Cotterill S. Synthesis of data from trials of interventions designed to change health behaviour; a case study. *Authorea*. 2021.
18. Higgins JPT, López-López JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM. Synthesising quantitative evidence in systematic reviews of complex health interventions. *2019*;4:e000858.
19. Harrison S, Jones HE, Martin RM, Lewis SJ, Higgins JPT. The albatross plot: A novel graphical tool for presenting results of diversely reported studies in a systematic review. *Research Synthesis Methods*. 2017;8(3):281-9.
20. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD. Audit and feedback: effects on professional practice and healthcare outcomes. *2012*;6.
21. Smith SM, Dworkin RH, Turk DC, McDermott MP, Eccleston C, Farrar JT, et al. Interpretation of chronic pain clinical trial outcomes: IMMPACT recommended considerations. *Pain*. 2020;161(11):2446-61.
22. Horrobin DF. Are large clinical trials in rapidly lethal diseases usually unethical? *Lancet*. 2003;361(9358):695-7.
23. Cook JA, Julious SA, Sones W, Hampson LV, Hewitt C, Berlin JA, et al. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Trials*. 2018;19(1):606.
24. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet*. 2005;365(9467):1348-53.
25. Campbell MK, Grimshaw JM, Elbourne DR. Intracluster correlation coefficients in cluster randomized trials: empirical insights into how should they be reported. *BMC Medical Research Methodology*. 2004;4(1):9.
26. Ukoumunne OC, Gulliford MC, Chinn S. A note on the use of the variance inflation factor for determining sample size in cluster randomized trials. *Journal of the Royal Statistical Society: Series D (The Statistician)*. 2002;51(4):479-84.
27. Hemming K, Kasza J, Hooper R, Forbes A, Taljaard M. A tutorial on sample size calculation for multiple-period cluster randomized parallel, cross-over and stepped-wedge trials using the Shiny CRT Calculator. *International Journal of Epidemiology*. 2020;49(3):979-95.
28. Batistatou E, Roberts C, Roberts S. Sample Size and Power Calculations for Trials and Quasi-Experimental Studies with Clustering. *The Stata Journal*. 2014;14(1):159-75.
29. Eldridge SM, Costelloe CE, Kahan BC, Lancaster GA, Kerry SM. How big should the pilot study for my cluster randomised trial be? *Stat Methods Med Res*. 2016;25(3):1039-56.
30. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004;57(8):785-94.
31. Thompson DM, Fernald DH, Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health projects. *Ann Fam Med*. 2012;10(3):235-40.
32. Health Services Research Unit Research Tools: University of Aberdeen; [Available from: <https://www.abdn.ac.uk/hsrcu/what-we-do/tools/index.php#panel177>].
33. Forster A, Dickerson J, Young J, Patel A, Kalra L, Nixon J, et al. A structured training programme for caregivers of inpatients after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness analysis. *Lancet*. 2013;382(9910):2069-76.
34. Cooper B, Kinsella GJ, Picton C. Development and initial validation of a family appraisal of caregiving questionnaire for palliative care. *Psychooncology*. 2006;15(7):613-22.
35. Aoun SM, Grande G, Howting D, Deas K, Toye C, Troeung L, et al. The Impact of the Carer Support Needs Assessment Tool (CSNAT) in Community Palliative Care Using a Stepped Wedge Cluster Trial. *PLOS ONE*. 2015;10(4):e0123012.

36. Walters SJ, Bonacho dos Anjos Henriques-Cadby I, Bortolami O, Flight L, Hind D, Jacques RM, et al. Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. *BMJ Open*. 2017;7(3):e015276.
37. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44(3):1051-67.
38. 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(5):1292-300.
39. Hemming K, Marsh J. A menu-driven facility for sample-size calculations in cluster randomized controlled trials. *Stata Journal*. 2013;13(1):114-35.
40. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. *Trials*. 2015;16(1):405.
41. 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. *International journal of surgery (London, England)*. 2012;10(1):28-55.
42. Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery TR, et al. Targeted versus Universal Decolonization to Prevent ICU Infection. *New England Journal of Medicine*. 2013;368(24):2255-65.
43. Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Medical Research Methodology*. 2005;5(1):10.
44. Giraudeau B, Ravaud P. Preventing Bias in Cluster Randomised Trials. *PLOS Medicine*. 2009;6(5):e1000065.
45. Yang R, Carter BL, Gums TH, Gryzlak BM, Xu Y, Levy BT. Selection bias and subject refusal in a cluster-randomized controlled trial. *BMC Medical Research Methodology*. 2017;17(1):94.
46. Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ*. 2003;327(7418):785.
47. Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*. 2008;336(7649):876.
48. Brierley G, Brabyn S, Torgerson D, Watson J. Bias in recruitment to cluster randomized trials: a review of recent publications. *J Eval Clin Pract*. 2012;18(4):878-86.
49. Bolzern J, Mnyama N, Bosanquet K, Torgerson D. Comparing evidence of selection bias between cluster-randomised and individually randomised controlled trials: a systematic review and meta-analysis. *The Lancet (British edition)*. 2018;392:S21-S.
50. Wade J, Elliott D, Avery KNL, Gaunt D, Young GJ, Barnes R, et al. Informed consent in randomised controlled trials: development and preliminary evaluation of a measure of Participatory and Informed Consent (PIC). *Trials*. 2017;18(1):327.
51. Rebers S, Aaronson NK, van Leeuwen FE, Schmidt MK. Exceptions to the rule of informed consent for research with an intervention. *BMC Med Ethics*. 2016;17:9-.
52. Taljaard M, McRae AD, Weijer C, Bennett C, Dixon S, Taleban J, et al. Inadequate reporting of research ethics review and informed consent in cluster randomised trials: review of random sample of published trials. *BMJ*. 2011;342:d2496.
53. Eldridge SM, Ashby D, Feder GS. Informed patient consent to participation in cluster randomized trials: an empirical exploration of trials in primary care. *Clinical trials (London, England)*. 2005;2(2):91-8.
54. Weijer C, Grimshaw JM, Taljaard M, Binik A, Boruch R, Brehaut JC, et al. Ethical issues posed by cluster randomized trials in health research. *Trials*. 2011;12(1):100.

55. Nix HP, Weijer C, Brehaut JC, Forster D, Goldstein CE, Taljaard M. Informed consent in cluster randomised trials: a guide for the perplexed. *BMJ Open*. 2021;11(9):e054213.
56. Weijer C, Taljaard M. The ethics of cluster randomized trials: response to a proposal for revision of the Ottawa Statement. *J Clin Epidemiol*. 2019;116:140-5.
57. Gallo A, Weijer C, White A, Grimshaw JM, Boruch R, Brehaut JC, et al. What is the role and authority of gatekeepers in cluster randomized trials in health research? *Trials*. 2012;13(1):116.
58. 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 medicine*. 2012;9(11):e1001346-e.
59. van der Graaf R, Koffijberg H, Grobbee DE, de Hoop E, Moons KGM, van Thiel GJM, et al. The ethics of cluster-randomized trials requires further evaluation: a refinement of the Ottawa Statement. *Journal of Clinical Epidemiology*. 2015;68(9):1108-14.
60. Goldstein CE, Weijer C, Taljaard M, Al-Jaishi AA, Basile E, Brehaut J, et al. Ethical Issues in Pragmatic Cluster-Randomized Trials in Dialysis Facilities. *American Journal of Kidney Diseases*. 2019;74(5):659-66.
61. Eldridge S, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*. 2009;339:b4006.
62. Giraudeau B, Caille A, Le Gouge A, Ravaud P. Participant Informed Consent in Cluster Randomized Trials: Review. *PLOS ONE*. 2012;7(7):e40436.
63. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*. 1999;319(7211):670.
64. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*. 2007;26(1):20-36.
65. Leyrat C, Caille A, Donner A, Giraudeau B. Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine*. 2013;32(19):3357-72.
66. Taft AJ, Small R, Hegarty KL, Watson LF, Gold L, Lumley JA. Mothers' AdvocateS In the Community (MOSAIC)- non-professional mentor support to reduce intimate partner violence and depression in mothers: a cluster randomised trial in primary care. *BMC Public Health*. 2011;11(1):178.
67. Twisk JWR, de Vente W. The analysis of randomised controlled trial data with more than one follow-up measurement. A comparison between different approaches. *European Journal of Epidemiology*. 2008;23(10):655-60.
68. Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*. 2014;15(1):139.
69. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice*. 2000;17(2):192-6.
70. Kerry SM, Bland JM. Analysis of a trial randomised in clusters. *Bmj*. 1998;316(7124):54.
71. Donner A, Klar N. Methods for Comparing Event Rates in Intervention Studies When the Unit of Allocation is a Cluster. *American Journal of Epidemiology*. 1994;140(3):279-89.
72. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. *Cochrane handbook for systematic reviews of interventions*. Second edition. ed. Hoboken, NJ: Wiley-Blackwell; 2019.
73. Moerbeek M. Cluster Randomized Trials: Design and Analysis. In: Pham H, editor. *Springer Handbook of Engineering Statistics*. London: Springer London; 2006. p. 705-18.

74. Bell ML, Rabe BA. The mixed model for repeated measures for cluster randomized trials: a simulation study investigating bias and type I error with missing continuous data. *Trials*. 2020;21(1):148.
75. Huang S, Fiero MH, Bell ML. Generalized estimating equations in cluster randomized trials with a small number of clusters: Review of practice and simulation study. *Clinical Trials*. 2016;13(4):445-9.
76. Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, et al. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials*. 2016;17(1):438.
77. Bellamy SL, Gibberd R, Hancock L, Howley P, Kennedy B, Klar N, et al. Analysis of dichotomous outcome data for community intervention studies. *Stat Methods Med Res*. 2000;9(2):135-59.
78. Ukoumunne OC, Carlin JB, Gulliford MC. A simulation study of odds ratio estimation for binary outcomes from cluster randomized trials. *Stat Med*. 2007;26(18):3415-28.
79. Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used? *International Journal of Epidemiology*. 2018;47(1):321-31.
80. Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Statistics in Medicine*. 2007;26(19):3550-65.
81. Fiero MH, Huang S, Oren E, Bell ML. Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*. 2016;17(1):72.
82. Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American journal of public health*. 1995;85(10):1378-83.
83. Eldridge S, Kerry SM. *A practical guide to cluster randomised trials in health services research*. Chichester, West Sussex: John Wiley & Sons; 2012.
84. Gamble C, Krishan A, Stocken D, Lewis S, Juszcak E, Doré C, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *Jama*. 2017;318(23):2337-43.
85. Marchenko YV. Estimating variance components in Stata. *Stata Journal*. 2006;6(1):1-21.
86. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. *PLOS ONE*. 2016;11(3):e0150205.
87. Hooper R, Forbes A, Hemming K, Takeda A, Beresford L. Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ*. 2018;360:k1121.
88. Copas AJ, Hooper R. Cluster randomised trials with different numbers of measurements at baseline and endline: Sample size and optimal allocation. *Clinical Trials*. 2019;17(1):69-76.
89. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ : British Medical Journal*. 2015;350:h391.
90. Elsheikh MA, Moriyama M, Rahman MM, Kako M, El-Monshed AH, Zoromba M, et al. Effect of a tailored multidimensional intervention on the care burden among family caregivers of stroke survivors: study protocol for a randomised controlled trial. *BMJ open*. 2020;10(12):e041637.
91. Kontou E, Thomas SA, Copley C, Fisher R, Golding-Day MR, Walker MF. A Biopsychosocial Intervention for Stroke Carers (BISC): development and description of the intervention. *Health Psychology and Behavioral Medicine*. 2022;10(1):92-103.

92. Krnic Martinic M, Pieper D, Glatt A, Puljak L. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. *BMC Medical Research Methodology*. 2019;19(1):203.
93. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986;7(3):177-88.
94. Evans D. Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*. 2003;12(1):77-84.
95. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. *Evidence Based Medicine*. 2016;21(4):125.
96. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clinical research ed)*. 1986;292(6522):746.
97. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010;1(2):97-111.
98. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics*. 1992;48(2):577-85.
99. Laopaiboon M. Meta-analyses involving cluster randomization trials: a review of published literature in health care. *Statistical Methods in Medical Research*. 2003;12(6):515-30.
100. Richardson M, Garner P, Donegan S. Cluster Randomised Trials in Cochrane Reviews: Evaluation of Methodological and Reporting Practice. *PloS one*. 2016;11(3):e0151818-e.
101. Hemming K, Taljaard M, Weijer C, Forbes AB. Use of multiple period, cluster randomised, crossover trial designs for comparative effectiveness research. *BMJ*. 2020;371:m3800.
102. McIntyre L, Taljaard M, McArdle T, Fox-Robichaud A, English SW, Martin C, et al. FLUID trial: a protocol for a hospital-wide open-label cluster crossover pragmatic comparative effectiveness randomised pilot trial. *BMJ Open*. 2018;8(8):e022780.
103. Morgan KE, Forbes AB, Keogh RH, Jairath V, Kahan BC. Choosing appropriate analysis methods for cluster randomised cross-over trials with a binary outcome. *Statistics in medicine*. 2017;36(2):318-33.
104. Grantham KL, Kasza J, Heritier S, Hemming K, Forbes AB. Accounting for a decaying correlation structure in cluster randomized trials with continuous recruitment. *Statistics in medicine*. 2019;38(11):1918-34.
105. Li F, Forbes AB, Turner EL, Preisser JS. Power and sample size requirements for GEE analyses of cluster randomized crossover trials. *Statistics in medicine*. 2019;38(4):636-49.
106. Turner RM, White IR, Croudace T. Analysis of cluster randomized cross-over trial data: a comparison of methods. *Statistics in medicine*. 2007;26(2):274-89.
107. Hooper R, Teerenstra S, de Hoop E, Eldridge S. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*. 2016;35(26):4718-28.
108. Hooper R, Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ : British Medical Journal*. 2015;350:h2925.
109. Climo MW, Yokoe DS, Warren DK, Perl TM, Bolon M, Herwaldt LA, et al. Effect of Daily Chlorhexidine Bathing on Hospital-Acquired Infection. *New England Journal of Medicine*. 2013;368(6):533-42.
110. Lewis SR, Butler AR, Evans DJW, Alderson P, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. *Cochrane Database of Systematic Reviews*. 2016(6).

111. Bleasdale SC, Trick WE, Gonzalez IM, Lyles RD, Hayden MK, Weinstein RA. Effectiveness of chlorhexidine bathing to reduce catheter-associated bloodstream infections in medical intensive care unit patients. *Arch Intern Med*. 2007;167(19):2073-9.
112. Noto MJ, Domenico HJ, Byrne DW, Talbot T, Rice TW, Bernard GR, et al. Chlorhexidine Bathing and Health Care–Associated Infections: A Randomized Clinical Trial. *JAMA*. 2015;313(4):369-78.
113. Boonyasiri A, Thaisiam P, Permpikul C, Judaeng T, Suiwongsa B, Apiradeewajeset N, et al. Effectiveness of Chlorhexidine Wipes for the Prevention of Multidrug-Resistant Bacterial Colonization and Hospital-Acquired Infections in Intensive Care Unit Patients: A Randomized Trial in Thailand. *Infection Control & Hospital Epidemiology*. 2016;37(3):245-53.
114. Camus C, Sebillé V, Legras A, Garo B, Renault A, Le Corre P, et al. Mupirocin/chlorhexidine to prevent methicillin-resistant *Staphylococcus aureus* infections: post hoc analysis of a placebo-controlled, randomized trial using mupirocin/chlorhexidine and polymyxin/tobramycin for the prevention of acquired infections in intubated patients. *Infection*. 2014;42(3):493-502.
115. Swan JT, Ashton CM, Bui LN, Pham VP, Shirkey BA, Blackshear JE, et al. Effect of Chlorhexidine Bathing Every Other Day on Prevention of Hospital-Acquired Infections in the Surgical ICU: A Single-Center, Randomized Controlled Trial*. *Critical Care Medicine*. 2016;44(10).
116. Pallotto C, Fiorio M, De Angelis V, Ripoli A, Franciosini E, Quondam Girolamo L, et al. Daily bathing with 4% chlorhexidine gluconate in intensive care settings: a randomized controlled trial. *Clin Microbiol Infect*. 2019;25(6):705-10.
117. Milstone AM, Elward A, Song X, Zerr DM, Orscheln R, Speck K, et al. Daily chlorhexidine bathing to reduce bacteraemia in critically ill children: a multicentre, cluster-randomised, crossover trial. *Lancet*. 2013;381(9872):1099-106.
118. Campbell M, Grimshaw J, Steen N. Sample Size Calculations for Cluster Randomised Trials. *Journal of Health Services Research & Policy*. 2000;5(1):12-6.
119. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Stat Med*. 2002;21(19):2971-80.
120. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Statistical methods in medical research*. 2001;10(5):325-38.
121. Burch J, Weller C. What are the effects of chlorhexidine bathing for preventing hospital-acquired infection in people admitted to intensive care units (ICUs)? 2019 [Available from: <https://www.cochranelibrary.com/cca/doi/10.1002/cca.2716/full>].
122. Bero LA. Cochrane Special Collections. Coronavirus (COVID-19): infection control and prevention measures: Cochrane Collaboration; 2020 [updated 18 Jan 2022. Available from: <https://www.cochranelibrary.com/collections/doi/SC000040/full>].
123. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby J. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. 2021;374:n2061.
124. Cotterill S, Powell R, Rhodes S, Brown B, Roberts J, Tang MY, et al. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. *Systematic reviews*. 2019;8(1):176.
125. Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. *Evidence Based Mental Health*. 2016;19(1):16.
126. Petticrew M, Rehfuess E, Noyes J, Higgins JP, Mayhew A, Pantoja T, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol*. 2013;66(11):1230-43.
127. Squires JE, Valentine JC, Grimshaw JM. Systematic reviews of complex interventions: framing the review question. 2013;66:1215-22.

128. Petticrew M, Anderson L, Elder R, Grimshaw J, Hopkins D, Hahn R, et al. Complex interventions and their implications for systematic reviews: a pragmatic approach. *Journal of Clinical Epidemiology*. 2013;66(11):1209-14.
129. Datta J, Petticrew M. Challenges to evaluating complex interventions: a content analysis of published papers. *BMC Public Health*. 2013;13(1):568.
130. Harrison S, Jones HE, Martin RM, Lewis SJ, Higgins JPT. The albatross plot: A novel graphical tool for presenting results of diversely reported studies in a systematic review. *Res Synth Methods*. 2017;8(3):281-9.
131. Ogilvie D, Fayter D, Petticrew M, Sowden A, Thomas S, Whitehead M, et al. The harvest plot: A method for synthesising evidence about the differential effects of interventions. *BMC Med Res Methodol*. 2008;8(1):8.
132. Higgins JPT, López-López JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM, et al. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health*. 2019;4(Suppl 1):e000858.
133. Kahwati L, Jacobs S, Kane H, Lewis M, Viswanathan M, Golin CE. Using qualitative comparative analysis in a systematic review of a complex intervention. *Systematic reviews*. 2016;5(1):82.
134. Anderson LM, Oliver SR, Michie S, Rehfues E, Noyes J, Shemilt I. Investigating complexity in systematic reviews of interventions by using a spectrum of methods. *J Clin Epidemiol*. 2013;66(11):1223-9.
135. Noyes J, Booth A, Moore G, Flemming K, Tunçalp Ö, Shakibazadeh E. Synthesising quantitative and qualitative evidence to inform guidelines on complex interventions: clarifying the purposes, designs and outlining some methods. *BMJ Global Health*. 2019;4(Suppl 1):e000893.
136. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol*. 2008;8:45.
137. Anderson R, Hardwick R, Pearson M, Byng R. Using Realist Approaches to Explain the Costs and Cost-Effectiveness of Programmes. In: Emme N, Greenhalgh J, Manzano A, Monaghan M, Dalkin S, editors. *Doing Realist Research*. London: SAGE Publications Ltd; 2018.
138. Bor J, Moscoe E, Mutevedzi P, Newell ML, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. 2014;25:729-37.
139. Dias S, Caldwell DM. Network meta-analysis explained. *Archives of Disease in Childhood - Fetal and Neonatal Edition*. 2019;104(1):F8.
140. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-73.
141. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. 2013;46:81-95.
142. Guldborg T, Vedsted P, Kristensen J, Lauritzen T. Improved quality of Type 2 diabetes care following electronic feedback of treatment status to general practitioners: a cluster randomized controlled trial. 2011;28:325-32.
143. Freeman SC, Scott NW, Powell R, Johnston M, Sutton AJ, Cooper NJ. Component network meta-analysis identifies the most effective components of psychological preparation for adults undergoing surgery under general anesthesia. *Journal of Clinical Epidemiology*. 2018;98:105-16.
144. Boet S, Bryson GL, Taljaard M, Pigford A-A, Mclsaac DI, Brehaut J, et al. Effect of audit and feedback on physicians' intraoperative temperature management and patient outcomes: a three-arm cluster randomized-controlled trial comparing benchmarked and

- ranked feedback. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*. 2018;65(11):1196-209.
145. Lakshminarayan K, Borbas C, McLaughlin B, Morris NE, Vazquez G, Luepker RV, et al. A cluster-randomized trial to improve stroke care in hospitals. *Neurology*. 2010;74(20):1634.
146. Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, et al. Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices: A Randomized Clinical Trial. *JAMA*. 2016;315(6):562-70.
147. Hemming K, Taljaard M. Estimands in cluster trials: thinking carefully about the target of inference and the consequences for analysis choice. *International Journal of Epidemiology*. 2022:dyac174.
148. 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.
149. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ (Clinical research ed)*. 2019;364:k4817-k.
150. Shi C, Dumville JC, Cullum N, Rhodes S, McInnes E, Goh EL, et al. Beds, overlays and mattresses for preventing and treating pressure ulcers: an overview of Cochrane Reviews and network meta-analysis. *Cochrane Database of Systematic Reviews*. 2021(8).
151. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*. 2019;10(1):83-98.
152. Doi SAR, Furuya-Kanamori L, Thalib L, Barendregt JJ. Meta-analysis in evidence-based healthcare: a paradigm shift away from random effects is overdue. *JBIC Evidence Implementation*. 2017;15(4).
153. Poole C, Greenland S. Random-Effects Meta-Analyses Are Not Always Conservative. *American Journal of Epidemiology*. 1999;150(5):469-75.
154. Nikolakopoulou A, Mavridis D, Salanti G. Demystifying fixed and random effects meta-analysis. *Evidence Based Mental Health*. 2014;17(2):53.
155. Senn S. The Many Modes of Meta. *Drug information journal : DIJ / Drug Information Association*. 2000;34(2):535-49.

MY SUBMITTED PAPERS

6.1 PAPER1


Patchwood E, Rothwell K, Rhodes S, Batistatou E, Woodward-Nutt K, Lau Y-S, Grande G, Ewing G, Bowen A. Organising Support for Carers of Stroke Survivors (OSCARSS): study protocol for a cluster randomised controlled trial, including health economic analysis. *Trials*. 2019;20(1):19. <https://doi.org/10.1186/s13063-018-3104-7>

STUDY PROTOCOL

Open Access



Organising Support for Carers of Stroke Survivors (OSCARSS): study protocol for a cluster randomised controlled trial, including health economic analysis

Emma Patchwood^{1,2*} , Katy Rothwell¹, Sarah Rhodes^{1,3}, Evridiki Batistatou^{1,3}, Kate Woodward-Nutt¹, Yiu-Shing Lau⁴, Gunn Grande^{1,5}, Gail Ewing⁶ and Audrey Bowen^{1,2}

Abstract

Background: Stroke often results in chronic disability, with partners and family members taking on the role of informal caregiver. There is considerable uncertainty regarding how best to identify and address carers' needs. The Carer Support Needs Assessment Tool (CSNAT) is a carer-led approach to individualised assessment and support for caregiving that may be beneficial in palliative care contexts. CSNAT includes an implementation toolkit. Through collaboration, including with service users, we adapted CSNAT for stroke and for use in a UK stroke specialist organisation providing long-term support. The main aims of OSCARSS are to investigate the clinical and cost-effectiveness of CSNAT-Stroke relative to current practice. This paper focuses on the trial protocol, with the embedded process evaluation reported separately.

Methods: Longitudinal, multi-site, pragmatic, cluster randomised controlled trial with a health economic analysis. Clusters are UK services randomised to CSNAT-Stroke intervention or usual care, stratified by size of service. Eligible carer participants are: adults aged > 18 years; able to communicate in English; referred to participating clusters; and seen face-to-face at least once by the provider, for support. The 'date seen' for initial support denotes the start of intervention (or control) and carers are referred to the research team after this for study recruitment. Primary outcome is caregiver strain (FACQ - Strain) at three months after 'date seen'. Secondary outcomes include: caregiver distress; positive caregiving appraisals (both FACQ subscales); Pound Carer Satisfaction with Services; mood (HADs); and health (EQ-5D5L) at three months. All outcomes are followed up at six months. Health economic analyses will use additional data on caregiver health service utilisation and informal care provision.

(Continued on next page)

* Correspondence: emma.patchwood@manchester.ac.uk

¹National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care Greater Manchester (CLAHRC GM), Salford Royal Foundation NHS Trust, Salford, UK

²Division of Neuroscience and Experimental Psychology, School of Biological Sciences, University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK

Full list of author information is available at the end of the article



(Continued from previous page)

Discussion: OSCARSS is open to recruitment at the time of article submission. Study findings will allow us to evaluate the clinical and cost-effectiveness of the CSNAT-Stroke intervention, directed at improving outcomes for informal carers of stroke survivors. Trial findings will be interpreted in the context of our embedded process evaluation including qualitative interviews with those who received and provided services as well as data on treatment fidelity. OSCARSS will contribute to knowledge of the unmet needs of informal stroke caregivers and inform future stroke service development.

Trial registration: ISRCTN Registry, [ISRCTN58414120](https://www.isrctn.com/ISRCTN58414120). Registered on 26 July 2016.

Keywords: Cluster randomised controlled trial, Informal caregivers, Carers, Stroke, Complex intervention, Health service; service user involvement; health economics; qualitative interviews

Background

Stroke causes a greater range of disabilities than any other chronic condition in the UK [1]. Stroke survivors experience loss of abilities and independence and express concerns about how their condition impacts their partners and family members, who often take on the role of informal caregiver to support personal care and daily living [2, 3]. In the UK alone, informal caregivers for stroke provide care worth up to £2.5 billion per year [4, 5]. This can come at a great personal cost to informal carers, threatening their physical health, connection with family and social networks, finances and emotional wellbeing [6–9].

Identifying and addressing the needs of informal caregivers is a priority at a national level [10–12]. However, several Cochrane reviews highlight considerable uncertainty regarding how best to support stroke caregivers [13–15]. Research suggests that a ‘one-size fits all’ approach to assessment and support is not as beneficial as support that is most closely matched to individuals’ current and specific needs, priorities and preferences [16, 17].

The Carer Support Needs Assessment Tool (CSNAT) intervention [18] is a comprehensive carer-led approach to individualised assessment and support that was developed in the context of palliative care. It includes a staff training package and implementation toolkit. The CSNAT intervention appeared to reduce carer strain in a community palliative care context, when compared to a control of usual care in a before / after stepped wedge design [19]. It also appeared to improve carer psychological and physical health in bereavement in a UK stepped wedge trial [20]. In these pragmatic studies, no changes were made to other support services available for carers between control and intervention periods. Qualitative work with carers [21] and practitioners [22] suggested that CSNAT was highly valued by both groups and made best use of available resources and time when identifying and prioritising needs and supporting carers.

We adapted the original CSNAT intervention and training package for implementation in stroke practice, collectively named CSNAT-Stroke. The adaptation was carried out through close collaboration with carers and a UK stroke service provider organisation. A study-specific Research User Group (RUG) of individuals with experience of caring for a stroke survivor, was set up for OSCARSS and they support study development through regular meetings and representation on the Trial Management Group (TMG). They continue to input to study management while the trial is open to recruitment and thereafter will contribute to interpretation and dissemination of the findings.

In terms of service provider collaborators, a working group of senior Stroke Association staff and their Training and Development department collaborated in development of the staff training and implementation approach used in OSCARSS. The Stroke Association is a stroke specialist provider service with over 200 stroke support services throughout the UK. Services are organised flexibly to meet requirements of the local population; practice therefore varies across different services according to availability and preferences. Many services are embedded in hospitals and referrals for support are primarily received from the National Health Service (NHS) soon after the stroke event; although individuals can be referred – or self-refer – at any time after stroke. All OSCARSS research sites/clusters are drawn from Stroke Association services.

Trial aim and research questions

The primary aim of OSCARSS is to determine the effectiveness of the CSNAT-Stroke intervention for carers of stroke survivors, when compared to a usual care control. The primary research question is: does the intervention reduce caregiver strain (as measured by the strain subscale of the Family Appraisal of Caregiving Questionnaire (FACQ) [23]), when compared to control?

Secondary research questions address whether the intervention:

- reduces perceived caregiver distress (subscale of FACQ) [23];
- improves: carer perceptions of their health (EQ-5D-5 L) [24] and wellbeing (Hospital Anxiety and Depression Scale (HADS) [25]; positive caregiving appraisals (subscale of FACQ) [23]; and satisfaction with services (Pound Scale) [26];
- leads to less economic burden for carers and society (as measured by an adapted version of the Service Receipt Inventory [27] that records use of health services and informal care provision).

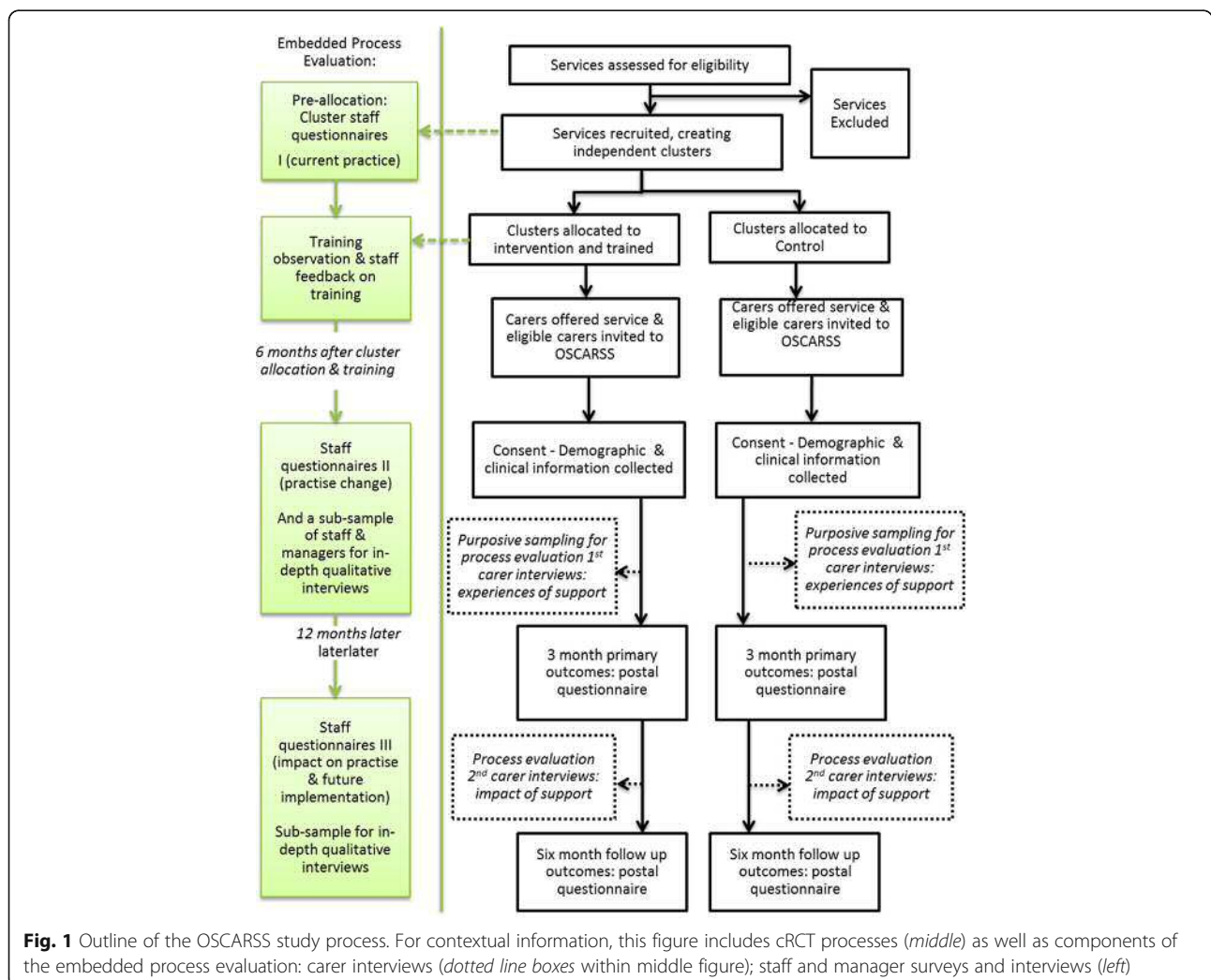
Methods

OSCARSS is a longitudinal, pragmatic multi-site cluster randomised controlled trial (cRCT) with a health economic analysis and nested process evaluation. Cluster randomisation is essential to avoid contamination as we

are evaluating delivery of an intervention within a service, sometimes by a team.

Not described in detail in this protocol is an embedded process evaluation that includes survey data, service delivery records and qualitative data. In brief, data collected from service providers (staff and managers) will explore intervention implementation and workforce behaviour change. Interviews with service recipients (carer research participants) will explore their experiences of support (intervention or control) and the types of support inputs identified and prioritised by them. Figure 1 shows these parallel components of OSCARSS but the cRCT and health economics are the focus of this paper. The process evaluation, which will be invaluable in providing the context for interpretation of the trial findings, will be described elsewhere.

A Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) checklist is provided as Additional file 1.



Trial clusters: inclusion criteria and randomisation

Eligible clusters are defined as UK stroke specialist provider services that:

- include face-to-face contact with carers (excluded are services that only provide telephone support);
- have capacity to engage in OSCARSS (excluded are services participating in any other stroke carer-related research);
- are likely to have at least five new client referrals per month. Clients include both stroke survivors (who are likely to have associated carers) and carers directly. This ensures that a new system of working can be operationalised and well-established and that research resources required for training and monitoring sites are justified;
- are independent of other clusters. If staff across services shared client caseloads, they would be aggregated to form one cluster to avoid the risk of between-group contamination. Conversely, individual staff within services could form independent clusters if they work independently, without sharing caseloads.

Clusters are recruited by the CLAHRC GM research team before randomisation, to ensure allocation concealment. Clusters are block randomised to intervention or control at the 'site' level with dichotomised stratification for 'size of service' (high / low -based on historic data about client caseloads) using random blocks. Neither the clusters nor the researchers know the block sizes when recruiting sites. The trial statistician is provided with an anonymised list of recruited clusters and randomises them using STATA programme, including the 'ralloc' add-on.

The research team is blinded to allocation as far as possible, but front-line team members could become unblinded when observing staff training (delivered after randomisation) or when supporting sites to engage in the study.

Carer research participants are blind to allocation; they receive support by the provider organisation in both arms of the trial, but the nature of support is different according to allocation to research intervention or control. Carers are not consenting to randomisation but to follow-up.

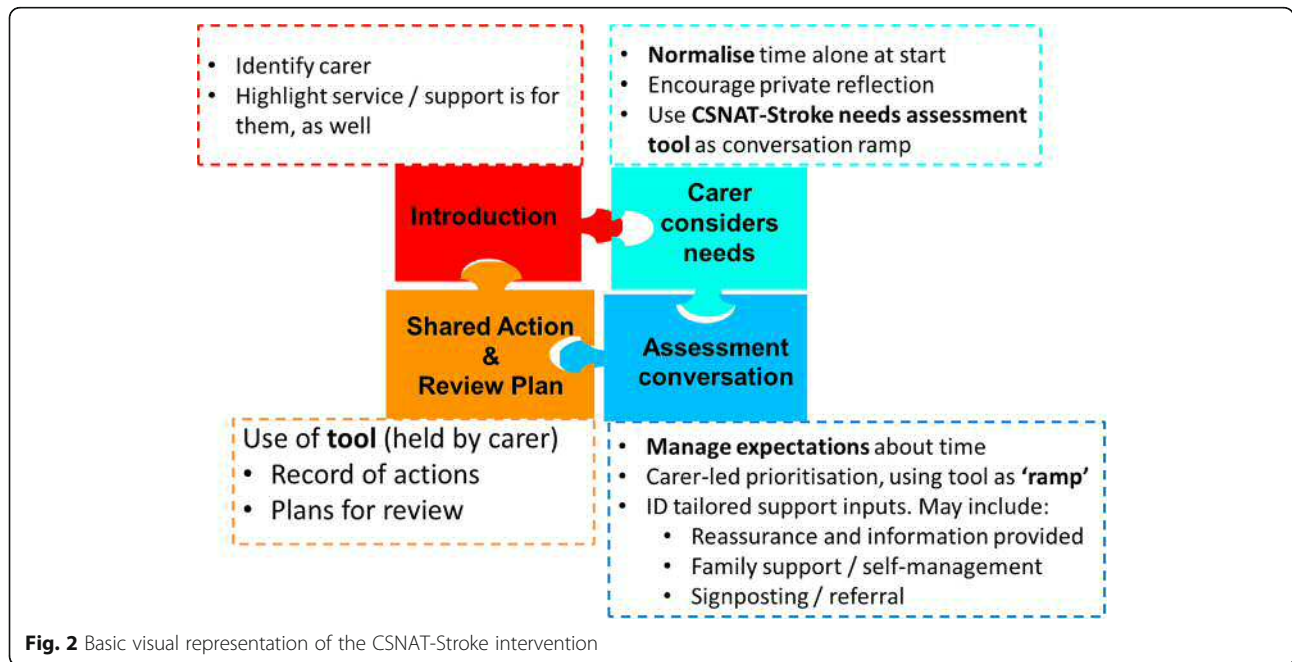
Intervention

The CSNAT-Stroke is the research intervention, described briefly here. All intervention materials, including training handbook and instructional videos, will be made available after the trial when treatment fidelity and adherence will be reported.

CSNAT-Stroke provides a structured, standardised approach to offering an evidence-based needs assessment for carers, which is distinct from the stroke survivor. It involves the use of a single-page assessment tool organised into broad domains of need and a written action plan for review. CSNAT-Stroke is predicated on staff behaviour change and follows a general process that can be flexibly applied whenever the staff member has contact with a carer. It is facilitated by a staff training and implementation package. CSNAT-Stroke promotes a carer-led, practitioner-facilitated approach to identifying and implementing support inputs that are directly derived from the needs assessment. Table 1 summarises the intervention process with Fig. 2 as a basic visual representation of the intervention. As described in our dissemination plan, we will report using the Template for

Table 1 Summary of CSNAT-Stroke intervention

Step	Description	Timing and duration	Mode
Introduction	Carers identified and assured that support is available	Point of referral to service; ≈ 5 min duration	Telephone or face-to-face (inpatient settings or home), depending on referral
Carers consider needs	CSNAT-Stroke needs assessment tool introduced. Carers encouraged to take independent time to consider and complete, indicating domains in which they need more support	At the point of contact; ≈ 5 min duration	Face-to-face (may be sent by post for follow-ups)
Assessment conversation	Using the CSNAT-Stroke tool as a 'conversation ramp', carers supported to prioritise the domains most important to them currently: to identify their individual needs within those domains and the type of supportive input they would find helpful. Support may be directly delivered by practitioner at this time (e.g. reassurance and information) but family support, signposting or referral to other agencies may also be included	During support contact. Duration dictated by time available; ≈ minimum 10 min, with 'set up' regarding time available to manage expectations	Typically face-to-face but possible by phone for follow-up contact
Shared action and review plan	Actions to address needs are recorded on a paper tool and, if appropriate, a plan is agreed regarding review / update on actions carried out	Following assessment conversation	Carers given hard copy action plan. Staff records in service database



Intervention Description and Replication (TIDieR) guidelines [28].

Carer research participants: inclusion criteria and recruitment processes

Adult (aged > 18 years) informal carers of stroke survivors are eligible if they:

- are referred to participating clusters;
- receive at least one face-to-face support contact (regardless of the resultant level of support / need, e.g. support may be a single face-to-face visit, with follow-up support by telephone);
- are able to communicate in English (facilitated by supportive communication techniques); and
- are 'active' in their caring role at the time of study entry, i.e. the stroke survivor being cared for is alive.

Following the first face-to-face contact, i.e. after support has been delivered according to intervention or control, staff provide a brief OSCARSS information leaflet and ask carers if they would like to be referred to the research team to find out more about potential study participation. Carers are told that the service is being evaluated, but they are not told about the randomised trial (blinding). The opportunity for study referral is offered, even if a carer does not go on to receive further support from the service. Carers can be given up to six weeks to decide about study referral.

If carers accept referral, their details are securely passed to the research team who make first contact by phone, introducing the study and providing full study information

by post to seek consent. This process ensures a clear separation between the research and the provision of support (either research intervention or control).

Informed consent is sought by researchers trained in Good Clinical Practice (GCP) and using approved documents for information and consent, which were co-designed with carers via the OSCARSS RUG. Participant information and consent materials are available on request from the authors and will be published at trial end. The right to refuse participation without giving reasons is respected and research participants remain free to withdraw at any time from the study without giving reasons and without prejudicing further support.

Data collection

The schedule of data collection and the outcome measures to be collected are shown in Fig. 3, the SPIRIT figure, with more detail below.

Carer self-report measures

Demographic and clinical characteristics, related to the carers and their cared-for stroke survivors, are collected at study entry, along with EQ-5D-5L [24]. These data are not strictly speaking baseline, as support (either intervention or control) has already been initiated at the point of study referral. Staff provide a 'date first seen' when referring carers to the study and this is considered the 'start date' for intervention or control.

Initial and follow-up outcomes are sought three and six months after 'start date', respectively. The time difference between study entry and outcomes collection will not necessarily be exactly three and six months,

	Enrolment	Allocation through cluster randomisation	Post-enrolment	Carer recruitment and data collection period (months)		
TIMEPOINT	-10 to 0 months	0	Month 3	Recruit 18 months (4 to 22)	3 month outcomes (7 to 25)	6 month outcomes (10 to 27#)
CLUSTER ENROLMENT:						
Cluster identification and eligibility screen	X					
Cluster Randomisation		X				
Cluster training			X			
CARER ENROLMENT				←————→		
INTERVENTIONS:						
<i>CSNAT-Stroke for carer support</i>				←————→		
<i>Standard practice carer support</i>				←————→		
ASSESSMENTS (postal questionnaire):						
<i>Demographic & clinical info</i>				X		
<i>EQ-5D 5L[†]</i>				X	X	X
<i>Caregiver Strain subscale of FACQ</i>					X*	X
<i>Caregiver distress subscale of FACQ</i>					X	X
<i>Positive caregiving appraisals subscale of FACQ</i>					X	X
<i>Pound Satisfaction with Stroke Services</i>					X	X
<i>HADS</i>					X	X
<i>Service Receipt Inventory (& time spent caring)[†]</i>					X	X
<p>*Primary outcome (caregiver strain subscale of FACQ at 3 month outcomes collection)</p> <p>[†] Health Economics related</p> <p># Due to overall study duration, data collection for six month (secondary) outcomes will end in month 27</p>						

Fig. 3 SPIRIT figure

since the recruitment process takes some time after referral (and, as above, referrals can be received up to six weeks from ‘start date’). In cases where carers request more time to make a decision about consent, study entry data and three-month outcomes can be collected simultaneously. To allow sufficient time for reminders and return post, initial outcomes can be returned any time up to 4.5 months from ‘start date’.

Follow-up outcome data are sought at six months and can be returned any time up to 7.5 months from ‘start date’. Due to the study end date, we can only collect six-month (secondary) outcomes up to month 27. A study database auto-generates all prompts for data collection and, to improve retention, phone calls engage participants before any postal packs are sent. Thank you notes also advise when participation is

complete and that a final report on results will be sent at study close (see ‘Dissemination plan’).

All measures are described below and are completed by carers via self-report postal questionnaire. Carers are given the option to complete over the phone with telephone support from a researcher:

- FACQ [23]: the caregiver strain (primary outcome at the three-month collection point) and caregiver distress subscales assess the negative impact of caring, while the positive appraisals subscale assess the positive impact of caring. The strain subscale of FACQ was used to successfully evaluate the effectiveness of the original CSNAT in a palliative care context [19] and is chosen as the primary clinical endpoint as it directly addresses the primary research aims. In addition, the OSCARSS RUG agreed that caregiver strain was most likely to be alleviated through this intervention and felt that, when compared to other candidate caregiver strain or burden scales, FACQ was more relatable and more likely to be completed through postal questionnaire;
- caregiver’s perceived quality of support and satisfaction with services will be assessed using the Pound Carer Satisfaction with Stroke Services Scale [26];
- carer wellbeing and health will be assessed using the HADS [25] and EQ-5D-5 L [24], respectively;
- for health economic analysis, an adapted version of the Service Receipt Inventory [27] will record use of health, social care and third sector services, as well as the amount and nature of informal care provision.

Service delivery records

Staff with access to the clusters’ in-house data management systems will securely provide study-specific data to the research team for consented carers. This will include: the dates, types and durations of support contacts delivered; and standardised entries from staff pertaining to needs identified and actions taken during support contacts. Support contacts include both direct and non-direct contact (e.g. liaison with external agents). Health economics analysis will include ‘service delivery costs’ for each consented carer, based on these data, by valuing support time using service provider full costs. As well as data specific to consented carers, the research team will be securely provided with fully anonymised service delivery records for all clients in participating services / clusters (intervention and control). These records will contain no personal client data but will include: the number, duration and type of contacts completed by coordinators; and categories of needs identified and actions completed. These data will support an economic understanding of whole service delivery across participating clusters (comparing intervention to control) and an exploration of how representative OSCARSS participants are of all cluster clients.

Sample size

The primary outcome is the Caregiver Strain subscale of the FACQ [23] at three months after intervention / control (see also Fig. 3). This subscale scale consists of eight questions, each worth a maximum of 5 points, and can be reported as a mean score per question (maximum score = 5.0) or total number of points (maximum score = 40 points). Cooper et al. [23] reported a mean (SD) of 3.13 (0.87) on this subscale on a study of 160 participants. In their trial to assess the impact of the CSNAT intervention in the palliative care setting, Aoun [19] reported a standardised effect size on the FACQ caregiver strain subscale of 0.348 which corresponds to a difference of 0.31 on the mean score. Based on empirical data from similar settings, we do not expect the intraclass correlation coefficient (ICC) to be > 0.05 (<https://www.abdn.ac.uk/hsru/what-we-do/tools>). In fact, TRACS, a cluster randomised trial which trained carers to provide care to stroke survivors [27], reported ICCs of 0.013 for caregiver burden.

Table 2 shows the sample sizes to achieve 80% power, assuming at least 16 active clusters per arm and SD = 0.9.

Our minimum target is 320 carers providing primary outcomes at three months. This would allow us 80% power to detect effect sizes of 0.31 or more for ICCs ≤ 0.01 , and effect sizes of ≥ 0.375 for ICCs of ≤ 0.05 . We assume a retention rate of 80% between consent and primary outcomes, which means we require a minimum of 400 consented carers.

An optimum sample size of 512 (640 consented carers) would allow us 80% power to detect effect sizes of ≥ 0.31 for ICCs ≤ 0.05 and would allow us to detect effect sizes of ≤ 0.25 for an ICC of 0.01. We would cease recruitment if we hit this figure before the planned recruitment end date.

Sample size calculations were carried out using the *clsampsi* function in STATA.

Statistical analysis

Adverse events

This study’s intervention is low risk, primarily involving staff behaviour change when supporting carers within their role. Serious adverse events (SAEs) are an inherent

Table 2 Sample size projections

ICC	Effect size = 0.31 unit change in mean score (2.5 points change on total score)
0	288
0.01	320
0.025	384
0.05	512
0.075	800

part of an active caregiving role (e.g. musculoskeletal injury; new medical problems or deterioration of existing medical problems, including depression). It is possible that these could lead to hospitalisation, prolongation of existing hospitalisation, disability / incapacity or death. As such, they are expected SAEs; there are no SAEs that we predict will be related to the research intervention. All adverse events (AEs) will be recorded. SAEs will be reported to the Research Ethics Committee (REC) within 15 days if the Chief Investigator believes they might be related to the research and unexpected.

Analysis, including economic evaluation

A full and detailed statistical analysis plan (SAP), including information on how any missing data will be managed, is included as an Additional file 2.

Analysis of the primary outcome comparing intervention and control at three months will be carried out on the basis of intervention to treat (ITT) and performed using a multilevel regression model, with a random intercept for 'site' to take into account clustering and a fixed covariate for 'intervention' along with adjustment using the following fixed individual level covariates: stroke severity of cared-for person (as rated by carer), time post-stroke, age of carer, health of carer at study entry (as indicated by self-reported pre-existing long-term health conditions) and the following cluster level covariates; size of service, pre-existing knowledge/experience of staff delivering support. By the design of this cluster randomised trial, recruitment of individual carers takes place after randomisation and therefore we are at risk of selection bias. We plan to adjust for baseline covariates in an attempt to control for any baseline imbalance. Similar analysis will be used for all numeric secondary outcome measures.

The mean number of carers per cluster, the mean number of support contacts per carer per cluster and the mean duration of contacts per carer per cluster will be compared between control and intervention groups using t-tests. We would not expect these variables to have appropriate distributions for analysis using a linear mixed model.

Sensitivity analyses will explore any potential bias in the analysis of the primary outcome measure and examine how robust the findings are:

- i. without adjustment for covariates;
 - ii. per protocol;
 - iii. combining three-month and six-month month data: using 'time' and 'time by group interaction' as fixed covariates, all available three-month and six-month data will be combined for the Caregiver Strain subscale of the FACQ. This will allow us to explore how caregiver strain changes over time
- iv. multiple imputation: using multiple imputation to replace missing values on the primary outcome measure using the following covariates: stroke severity of cared-for's stroke; time after stroke; age of carer; pre-stroke health of carer (as per Royston [29]);
 - v. excluding delayed responses: excluding any data from individuals who return their three-month outcome data later than 4.5 months after 'date seen' or six-month outcome data later than 7.5 months after 'date seen';
 - vi. removing carer dyads: where multiple carers of the same stroke survivor have provided outcome data; excluding data from the second and subsequent carers linked to the same stroke survivor.

Data relevant to the Health Economics analysis will include the Service Receipt Inventory, number of support contacts delivered per carer, informal care provision estimates and EQ-5D-5 L. Trial health economists will attach costs to questionnaire items and support contacts to allow a comparison across research intervention and control arms of the trial. Prognostically important variables such as carer health and demographics will be factored into an analysis comparing use of healthcare services, with severity of stroke survivor factored in to an analysis comparing time spent caring.

Data management and monitoring

All information collected is kept strictly confidential. Information will be held securely on paper in locked filing cabinets and electronically on encrypted servers. All data are anonymised as early as possible, with carers assigned a unique identifier as soon as they are entered into the database. If a participant withdraws consent at any time, their research data will remain on file and will be included in the final study analysis, unless otherwise requested. If a withdrawing participant agrees to receive a final report summarising the results of the study, their contact information will be held on file for these purposes and will be deleted once the final report is sent.

Standard Operating Procedures (SOP) for data entry processes ensure consensus in interpreting ambiguous data. The SOP also outlines data checking for quality and is available on request. Delegation logs determine which study staff are trained and assured to carry out specific tasks, including data entry.

The Research Team will form a Trial Management Group (TMG) and a Trial Steering Committee (TSC). The TSC will be chaired by and include independent members as well as key trial personnel. Data to be regularly monitored will include: individual level study-entry

demographic and clinical variables; and cluster level data related to referrals and recruitment. The TMG and TSC will consider recruitment and balance across the intervention and control arms throughout the study. After four months of carer recruitment, the TSC met to consider these data to make a recommendation as to whether the trial should be allowed to continue, continue with modification or be discontinued (they decided on the former). A TSC charter outlining roles and responsibilities is available on request.

Discussion

This paper describes the protocol for a novel trial exploring clinical and cost-effectiveness of a pragmatic intervention to support and empower informal carers of stroke survivors. It differs from TRACS [27] in that OSCARSS trains staff to support carers' own needs whereas the focus of TRACS was to train carers to perform the caring role. The OSCARSS intervention has been adapted from an existing approach used successfully in palliative care settings.

All aspects of the study have been designed in collaboration with key stakeholders, including carers themselves who form our study specific RUG, and stroke professionals who deliver the support. We believe this collaboration strengthens the study, including optimising recruitment processes and outcome measurement.

There are also some challenges to address, including the lack of baseline measures to explore change in outcomes, which the randomised design helps overcome. In terms of outcomes, the majority of our clinical endpoint data will be based on carer self-report, using measures that have been carefully selected through consultation with literature and co-development with service users. The intervention aims to provide individualised carer support and reduce the negative impact of caregiving, but our carer eligibility criteria are extremely inclusive and do not require diagnosis of depression, anxiety or similar. As such, hard clinical endpoints requiring professional assessment would be inappropriate in this pragmatic trial.

The decision to widen the time window for returning the postal questionnaire is a pragmatic one but may increase variability in when we measure outcomes. This will be adjusted for, as needed, with sensitivity analysis. Cluster randomisation is essential to avoid contamination as we are evaluating delivery of an intervention within a service but leads to potential for differential recruitment as allocation is known in advance of consent. Methods to overcome this have been outlined above and in the SAP but in addition, all cluster staff are given similar training with regards to recruitment and record keeping and are regularly engaged with by the research team and service managers to encourage consistent referrals to the study. Generalisability will be explored through

comparing characteristics of our sample to anonymised data related to national caseloads of the service provider.

This paper has focused on the cRCT and health economics, but it is strengthened by an embedded mixed-methods process evaluation to ensure a contextualised interpretation of our findings. The process evaluation will be described fully elsewhere but includes: implementation of the research into practice; sustainability of the research intervention; and the effect of research team on staff behaviour. The research intervention requires staff behaviour change and the pragmatic design leads to anticipated challenges exploring intervention fidelity, which the process evaluation will also help overcome. Semi-structured qualitative interviews will explore staff and carer experiences of delivering and receiving support, respectively. Interviews are completed with purposively sampled participants, considering demographic variables, arm allocation and geographical location. Interviews and focus groups will also be completed with service provider managers and senior leadership teams. The process evaluation is overseen by expert implementation and qualitative researchers who were not involved in the trial design.

Overall, OSCARSS will provide pragmatic data on future healthcare development for supporting carers of stroke survivors. Health economics components will allow exploration of costs with a view to providing a costed service specification to directly inform service improvements. The model for adapting and implementing the research intervention through collaboration could be applied to other health conditions and settings.

Dissemination plan

The findings from OSCARSS will be published in scientific journals using the following guidelines: Consolidated Standards of Reporting Trials (CONSORT) guidelines for cRCTs [30]; Template for Intervention Description and Replication (TIDieR) guidelines for intervention description [28]; and Consolidated criteria for reporting qualitative research (COREQ) guidance for qualitative research [31]. Trial findings will also be written up in accessible, lay-friendly language and disseminated to research participants and on the NIHR CLAHRC Greater Manchester website. A study-specific event to disseminate to all stakeholders will be held and we will disseminate to wider audiences through local, national and international conferences. Implementation activities will be finalised once the results are known.

Trial status

Clusters were randomised in September 2016 and trained in January 2017, when carer participant recruitment began. The first carer was enrolled on 17 January 2017. Recruitment is ongoing (at the time of journal submission) and will be completed by 31 July 2018.

Additional files

Additional file 1: SPIRIT 2013 Checklist. (DOC 141 kb)

Additional file 2: Statistical Analysis Plan. (DOCX 59 kb)

Abbreviations

AEs: Adverse events; CLAHRC: Collaboration for Leadership in Applied Health Research and Care; COREQ: Consolidated criteria for reporting qualitative research; CRCT: Cluster randomised controlled trial; CSNAT: Carer Support Needs Assessment Tool; FACQ: Family Appraisal of Caregiving Questionnaire; GCP: Good Clinical Practice; HADS: Hospital Anxiety and Depression Scale; ICC: Intraclass correlation coefficient; ISRCTN: International Standard Randomised Controlled Trial Number; MAHSC: Manchester Academic Health Sciences Centre; NHS: National Health Service; NIHR: National Institute for Health Research; OSCARSS: Organising Support for Carers of Stroke Survivors; REC: Research Ethics Committee; RUG: Research User Group; SAE: Serious adverse event; SAP: Statistical analysis plan; SD: Standard deviation; SOP: Standard Operating Procedure; SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials; TIDIER: Template for Intervention Description and Replication; TMG: Trial Management Group; TSC: Trial Steering Committee; UK: United Kingdom

Acknowledgements

We would like to acknowledge the dedication of all Stroke Association coordinators, managers and senior staff who participate in and support the project. Also, members of the CLAHRC GM team are involved in administration, recruitment, data collection, management and analysis: Zoe Ashton, Caroline O'Donnell, Rose Crees, Aneela McAvooy, Sam Wilkinson, Amy Woodhouse and Sandra Talbot. And of course, our Carer Research User Group members: Kelly Burke, Natalie Halford, Christine Halford, Ben Wright, Geoff Heathcote and Kath Purcell. And finally, an acknowledgment to all carer research participants for their time and valuable input to the study.

Funding

This project was supported by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Greater Manchester (NIHR CLAHRC GM), in partnership with the Stroke Association. The funder had no role in the design of the study, data collection and analysis, decision to publish or preparation of the manuscript. However, the project outlined in this article may be considered to be affiliated to the work of the NIHR CLAHRC GM. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or the Stroke Association. The authors are part funded by the Stroke Association and the NIHR CLAHRC Greater Manchester and several of them hold other NIHR or Stroke Association grants.

Availability of data and materials

Not applicable.

Trial sponsor

University of Manchester (ref: 16400). Contact: fmhsethicsapps@manchester.ac.uk or Lynne.K.Macrae@manchester.ac.uk

Authors' contributions

AB, KR and EP conceived the idea for this study. EP led on writing this paper. AB, KR, EP, GG, GE, SR, YSL, KWN and EB contributed to study design, acquisition of data and drafting of this paper and approved the final version. AB and EP are joint chief investigators. KWN is the trial manager and KR was the CLAHRC stroke programme manager. SR and EB are the trial statisticians. YSL is the health economist. GG and GE are the co-authors of the original CSNAT intervention. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethical approval has been obtained from the North West - Lancaster Research Ethics Committee (ref: 16/NW/0657) for the original protocol as well as all amendments. Written informed consent is obtained before carer or staff participant involvement in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests. The authors alone are responsible for the content and writing of the paper.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care Greater Manchester (CLAHRC GM), Salford Royal Foundation NHS Trust, Salford, UK. ²Division of Neuroscience and Experimental Psychology, School of Biological Sciences, University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK. ³Centre for Biostatistics, Institute of Population Health, University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK. ⁴Centre for Health Economics, Division of Population Health, Health Services Research & Primary Care, University of Manchester, Manchester, UK. ⁵Division of Nursing, Midwifery & Social Work, School of Health Sciences, University of Manchester, MAHSC, Manchester, UK. ⁶Centre for Family Research, University of Cambridge, Cambridge, UK.

Received: 14 June 2018 Accepted: 4 December 2018

Published online: 07 January 2019

References

- Adamson J, Beswick A, Ebrahim S. Is stroke the most common cause of disability? *J Stroke Cerebrovasc Dis.* 2004;13(4):171–7.
- McKevitt C, Redfern J, Mold F, Wolfe C. Qualitative studies of stroke: a systematic review. *Stroke.* 2004;35(6):1499–505.
- Patchick EL, Horne M, Woodward-Nutt K, Vail A, Bowen A. Development of a patient-centred, patient-reported outcome measure (PROM) for post-stroke cognitive rehabilitation: qualitative interviews with stroke survivors to inform design and content. *Health Expect.* 2015;18:2313–24.
- Luengo-Fernandez R, Leal J, Gray A, Petersen S, Rayner M. Cost of cardiovascular diseases in the United Kingdom. *Heart.* 2006;92(10):1384–9.
- Saka O, McGuire A, Wolfe C. Cost of stroke in the United Kingdom. *Age Ageing.* 2009;38(1):27–32.
- Draper P, Brocklehurst H. The impact of stroke on the well-being of the patient's spouse: an exploratory study. *J Clin Nurs.* 2007;16(2):264–71.
- Godwin KM, Ostwald SK, Cron SG, Wasserman J. Long-term health-related quality of life of stroke survivors and their spousal caregivers. *J Neurosci Nurs.* 2013;45(3):147–54.
- Byun E, Evans LK. Concept analysis of burden in caregivers of stroke survivors during the early poststroke period. *Clin Nurs Res.* 2015;24(5):468–86.
- Haley WE, Roth DL, Hovater M, Clay OJ. Long-term impact of stroke on family caregiver well-being: a population-based case-control study. *Neurology.* 2015;84(13):1323–9.
- Royal College of General Practitioners. RCGP Commissioning for Carers. London: RCGP; 2013. Available: <http://www.rcgp.org.uk/clinical-and-research/clinical-resources/~media/6C9C69A8CF64490594E464F6E11CC42.ashx>. Accessed 27 Apr 2018.
- Department of Health. Care Act. London: Department of Health; 2014. Available: <http://www.legislation.gov.uk/ukpga/2014/23/contents/enacted/data.htm>. Accessed 27 Apr 2018.
- Medical Directorate and Nursing Directorate. NHS England's Commitment for Carers. Leeds: NHS England; 2014. Available: <http://www.england.nhs.uk/wp-content/uploads/2014/05/commitment-to-carers-may14.pdf>. Accessed 27 Apr 2018.
- Ellis G, Mant J, Langhorne P, Dennis M, Winner S. Stroke liaison workers for stroke patients and carers: an individual patient data meta-analysis. *Cochrane Database Syst Rev.* 2010;5:CD005066. <https://doi.org/10.1002/14651858.CD005066.pub2>.
- Legg L, Quinn TJ, Mahmood F, Weir CJ, Tierney J, Stott DJ, et al. Non-pharmacological interventions for caregivers of stroke survivors. *Cochrane Database Syst Rev.* 2011;10:CD008179. <https://doi.org/10.1002/14651858.CD008179.pub2>.
- Forster A, Brown L, Smith J, House A, Knapp P, Wright JJ, et al. Information provision for stroke patients and their caregivers. *Cochrane Database Syst Rev.* 2012;11:CD001919. <https://doi.org/10.1002/14651858.CD001919.pub3>.

16. Sorensen S, Pinquart M, Duberstein P. How effective are interventions with caregivers? An updated meta-analysis. *Gerontologist*. 2002;42(3):356–72.
17. Cameron JI, Gignac MA. "Timing It Right": a conceptual framework for addressing the support needs of family caregivers to stroke survivors from the hospital to the home. *Patient Educ Couns*. 2008;70(3):305–14.
18. Ewing G, Grande G, National Association for Hospice at Home. Development of a Carer Support Needs Assessment Tool (CSNAT) for end-of-life care practice at home: a qualitative study. *Palliat Med*. 2013; 27(3):244–56.
19. Aoun SM, Grande G, Howting D, Deas K, Toye C, Troeung L, et al. The impact of the carer support needs assessment tool (CSNAT) in community palliative care using a stepped wedge cluster trial. *PLoS One*. 2015;10(4): e0123012.
20. Grande GE, Austin L, Ewing G, O'Leary N, Roberts C. Assessing the impact of a Carer Support Needs Assessment Tool (CSNAT) intervention in palliative home care: a stepped wedge cluster trial. *BMJ Support Palliat Care*. 2017;7: 326–34.
21. Aoun S, Deas K, Toye C, Ewing G, Grande G, Stajduhar K. Supporting family caregivers to identify their own needs in end-of-life care: Qualitative findings from a stepped wedge cluster trial. *Palliat Med*. 2015;29(6):508–17.
22. Ewing G, Austin L, Grande G. The role of the Carer Support Needs Assessment Tool in palliative home care: A qualitative study of practitioners' perspectives of its impact and mechanisms of action. *Palliat Med*. 2016; 30(4):392–400.
23. Cooper B, Kinsella GJ, Picton C. Development and initial validation of a family appraisal of caregiving questionnaire for palliative care. *Psychooncology*. 2006;15(7):613–22.
24. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–36.
25. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand*. 1983;67(6):361–70.
26. Pound P, Gompertz P, Ebrahim S. Development and results of a questionnaire to measure carer satisfaction after stroke. *J Epidemiol Community Health*. 1993;47(6):500–5.
27. Forster A, Dickerson J, Young J, Patel A, Kalra L, Nixon J, et al. A structured training programme for caregivers of inpatients after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness analysis. *Lancet*. 2013;382(9910):2069–76.
28. 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. *Gesundheitswesen*. 2016; 78(3):175–88.
29. Royston P. Multiple imputation of missing values: update of ice. *Stata J*. 2005;5(4):527.
30. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. Consort 2010 statement: extension to cluster randomised trials. *BMJ*. 2012; 345:e5661.
31. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care*. 2007;19(6):349–57.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions









6.2 PAPER 2

Patchwood E, Woodward-Nutt K, Rhodes SA, Batistatou E, Camacho E, Knowles S, Darley S, Grande G, Ewing G, Bowen A. Organising Support for Carers of Stroke Survivors (OSCARSS): a cluster randomised controlled trial with economic evaluation. *BMJ Open*.

2021;11(1):e038777. <http://dx.doi.org/10.1136/bmjopen-2020-038777>

BMJ Open Organising Support for Carers of Stroke Survivors (OSCARSS): a cluster randomised controlled trial with economic evaluation

Emma Patchwood ^{1,2}, Kate Woodward-Nutt,² Sarah A Rhodes,^{2,3} Evridiki Batistatou,^{2,3} Elizabeth Camacho ⁴, Sarah Knowles,^{2,5,6} Sarah Darley ^{2,5}, Gunn Grande ^{2,7}, Gail Ewing ^{2,8}, Audrey Bowen ^{1,2}

To cite: Patchwood E, Woodward-Nutt K, Rhodes SA, *et al*. Organising Support for Carers of Stroke Survivors (OSCARSS): a cluster randomised controlled trial with economic evaluation. *BMJ Open* 2021;**11**:e038777. doi:10.1136/bmjopen-2020-038777

► Prepublication history and supplemental material for this paper is available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2020-038777>).

Received 23 March 2020
Revised 10 December 2020
Accepted 14 December 2020



► <http://dx.doi.org/10.1136/bmjopen-2020-038777>



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Professor Audrey Bowen;
audrey.bowen@manchester.ac.uk

ABSTRACT

Objective Investigated clinical effectiveness and cost-effectiveness of a person-centred intervention for informal carers/caregivers of stroke survivors.

Design Pragmatic cluster randomised controlled trial (cRCT) with economic and process evaluation.

Setting Clusters were services, from a UK voluntary sector specialist provider, delivering support primarily in the homes of stroke survivors and informal carers.

Participants Adult carers in participating clusters were referred to the study by cluster staff following initial support contact.

Interventions Intervention was the Carer Support Needs Assessment Tool for Stroke: a staff-facilitated, carer-led approach to help identify, prioritise and address the specific support needs of carers. It required at least one face-to-face support contact dedicated to carers, with reviews as required. Control was usual care, which included carer support (unstructured and variable).

Outcome measures Participants provided study entry and self-reported outcome data by postal questionnaires, 3 and 6 months after first contact by cluster staff. Primary outcome: 3-month caregiver strain (Family Appraisal of Caregiving Questionnaire, FACQ). Secondary outcomes: FACQ subscales of caregiver distress and positive appraisals of caregiving, mood (Hospital Anxiety and Depression Scale) and satisfaction with stroke services (Pound). The economic evaluation included self-reported healthcare utilisation, intervention costs and EQ-5D-5L.

Randomisation and masking Clusters were recruited before randomisation to intervention or control, with stratification for size of service. Cluster staff could not be masked as training was required for participation. Carer research participants provided self-reported outcome data unaware of allocation; they consented to follow-up data collection only.

Results Between 1 February 2017 and 31 July 2018, 35 randomised clusters (18 intervention; 17 control) recruited 414 cRCT carers (208 intervention; 206 control). Study entry characteristics were well balanced. Primary outcome measure: intention-to-treat analysis for 84% retained participants (175 intervention; 174 control) found mean (SD) FACQ carer strain at 3 months to be 3.11 (0.87) in the control group compared with 3.03 (0.90) in the intervention group, adjusted mean difference of -0.04

Strengths and limitations of this study

- We successfully conducted the first adequately powered cluster randomised controlled trial of an approach to support informal carers of stroke survivors, but may have benefited from a feasibility trial to maximise intervention fidelity.
- We collaborated closely with service providers and previous service users to pragmatically tailor the intervention for implementation, including a staff-training package.
- The demographic profile of the sample was as expected for carers of stroke survivors but the sample lacked ethnic diversity and we may have benefited from seeking data beyond 6 months after support had been initiated.
- We highlight the feasibility of robust research with this population and signpost to suggestions from our nested process evaluation for improved implementation of person-centred care.

(95% CI -0.20 to 0.13). Secondary outcomes had similarly small differences and tight CIs. Sensitivity analyses suggested robust findings. Intervention fidelity was not achieved. Intervention-related group costs were marginally higher with no additional health benefit observed on EQ-5D-5L. No adverse events were related to the intervention. **Conclusions** The intervention was not fully implemented in this pragmatic trial. As delivered, it conferred no clinical benefits and is unlikely to be cost-effective compared with usual care from a stroke specialist provider organisation. It remains unclear how best to support carers of stroke survivors. To overcome the implementation challenges of person-centred care in carers' research and service development, staff training and organisational support would need to be enhanced.

Trial registration number ISRCTN58414120.

INTRODUCTION

Informal carers, providing unpaid support to family and friends with long-term health conditions, make an invaluable societal and

economic contribution. *BMJ* published the ‘unremitting burden on carers’ over 30 years ago,¹ but sadly carers’ own support needs are still often overlooked and being a caregiver often adversely affects health and well-being.²

Although countries such as the UK now mandate for the identification of carers’ support needs through the 2014 Care Act,³ less than one-third report receiving a statutory assessment.² One possible approach for comprehensive support is the Carer Support Needs Assessment Tool (CSNAT) intervention.⁴ The CSNAT intervention has multiple components including a comprehensive assessment tool integrated within a staged carer-led approach to individualised support. It was developed, implemented and tested in the context of palliative care with positive outcomes, including a significant reduction in caregiver strain as measured on the Family Appraisal of Caregiving Questionnaire (FACQ).^{5–9} We hypothesised that this intervention had the potential to support informal carers of people with long-term health conditions such as stroke, that causes a greater range of disabilities than any other in the UK.¹⁰ Recent systematic reviews and trials of carers of stroke survivors have highlighted the absence of a robustly proven support intervention.^{11–15}

In close collaboration with a study-specific carer advisory research group (see the Patient and public involvement section) and a UK stroke service provider organisation, we adapted the CSNAT intervention including a staff training and implementation package tailored to the provider organisation (see the Interventions and procedures section, and figure 1, table 1 and online supplemental table S1). This partnership was facilitated by the former National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Greater Manchester (NIHR CLAHRC GM, <https://www.clahrc-gm.nihr.ac.uk/>, now Applied Research Collaborations). The aim of the Organising Support for Carers of Stroke Survivors (OSCARSS) study was to determine the clinical and cost-effectiveness of the CSNAT-Stroke intervention for carers of stroke survivors, when compared with usual care. The primary hypothesis was that the

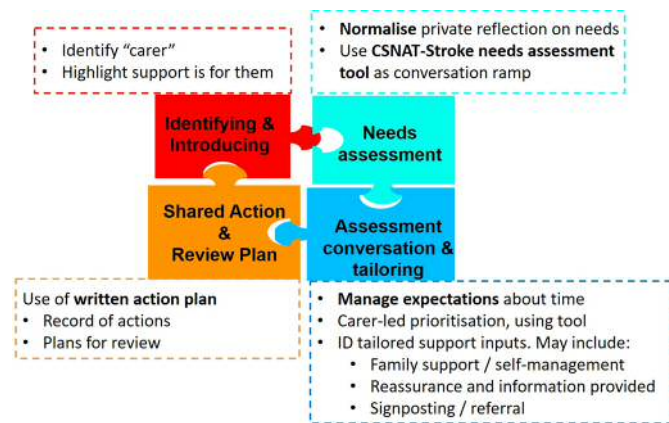


Figure 1 The adapted CSNAT-Stroke intervention as intended. CSNAT-Stroke, Carer Support Needs Assessment Tool for Stroke.

Table 1 Summary of key differences between intervention and usual care

Usual care	CSNAT-Stroke intervention
Focus primarily on stroke survivor	Focus specifically on carer survivor
No formal process with carers: varies across services	A standardised assessment and support process for carers services
Support carer if present	Make appointment to include carer
Usually see together with stroke survivor	Normalise seeing carer separately
If carer present: prompt question from practitioner about carer’s own needs	Carer-led assessment and prioritisation of needs using evidence-based assessment tool and staged person-centred approach
Review times vary	Carer-specific action and review plan

CSNAT-Stroke, Carer Support Needs Assessment Tool for Stroke.

adapted intervention would reduce caregiver strain when compared with usual care. Secondary hypotheses explored the impact on other aspects of the carer experience (eg, well-being and satisfaction with services), as well as its economic impact.

METHODS

Study design

OSCARSS was a longitudinal, pragmatic, national cluster randomised controlled trial (cRCT), underpinned by patient and public involvement from our study-specific carer advisory research group.¹⁶ Cluster randomisation was essential to avoid contamination. Clusters were drawn from services commissioned by the National Health Service (NHS) or local authorities, delivered by a UK voluntary sector stroke specialist organisation providing long-term support to stroke survivors and carers, including hospital and home visits. Eligible clusters were those with capacity for research participation and delivering support to carers in their own homes and a minimum of five new client (survivor or carer) referrals per month, based on historical service delivery records from a 9-month period before the study began.

This paper focuses on the RCT to explore the intervention’s clinical and cost-effectiveness. OSCARSS also included a mixed-methods embedded process evaluation to help understand intervention implementation and workforce adoption, described in detail elsewhere.¹⁷

Ethics approvals were obtained (see Ethics approval section) and the lead author (EP) affirms that this manuscript is an honest, accurate and transparent account of the study being reported. The study methods and design have been described in detail, with no major changes made to protocol.¹⁸

Patient and public involvement

A study-specific Research User Group (RUG) of 10 individuals with experience of caring for a stroke survivor was set up in December 2015, at the planning stages of OSCARSS. Through regular meetings (2015–2019) and representation on the Trial Management Group, the priorities, experiences and preferences of the RUG informed development of the research questions and the design, analysis/interpretation and dissemination of all components of the OSCARSS Study. The RUG supported authorship of an easy access report on the results of this study that has been sent to study participants (Dissemination Declaration).

The RUG advised on participant recruitment and were central in limiting the burden of participation for carers. The RUG also were key in supporting adaptation of the research intervention (CSNAT-Stroke) and staff training package, including role-playing videos of the intervention in practice. A video summarising their role in OSCARSS is available on the study website: <https://www.arc-gm.nihr.ac.uk/projects/oscarss>, and following GRIPP2 framework¹⁹ we have published a separate paper on the working practices and experiences of RUG members and the researchers who facilitated the group meetings.¹⁶

Participants

English-speaking informal carers of stroke survivors were eligible if they were over 18 years old and received at least one face-to-face support contact from participating cluster staff. Carers could be included at any time post-stroke event with any level of need or support requirements. We focused on those newly referred to the service as opposed to those using services for some time as core parts of the intervention included identification of carers. We aimed to recruit those individuals identified as ‘primary caregiver’, even when there may have been other informal carers involved.

Following the first face-to-face support contact (either intervention or control), eligible carers were invited by cluster staff to find out more about potential study participation and given up to 6 weeks to make a decision. Carers were assured that their decision on study participation would have no impact on the provision of ongoing support (either intervention or control). If carers accepted, their details were passed securely to the research team who provided full study information by post and sought informed written consent to participate. Procedures were also in place for consent to be taken by telephone. Researchers were in regular contact with all cluster staff and senior leadership to encourage fidelity with research procedures, including the consistent invitation to participate for all eligible carers.

Randomisation and masking

Details of the randomisation and masking were described in the protocol.¹⁸ Briefly, clusters were recruited (with consent of senior leadership and frontline staff within the provider organisation) by research staff before

randomisation to ensure allocation concealment at a cluster level. Clusters were block randomised to intervention or control, with stratification for size of service using random blocks of two (to ensure similar numbers of carers and clusters in each arm). The trial statistician performed the randomisation of all recruited clusters simultaneously using an anonymised list of cluster ID numbers and size of service data. The initial randomisation list produced allocations for 36 clusters, with a second randomisation list produced to allocate up to 16 clusters in the event of needing to replace clusters that dropped out or failed to recruit.

Cluster staff could not be masked as training was required to equip them to participate in the study. Training included participant recruitment and trial procedures (control and intervention arms of the trial) and the intervention (intervention arm only). The research team were masked to allocation as far as possible, although some team members could become unmasked during cluster staff training or support activities. Carer research participants provided self-report primary and secondary outcomes unaware of allocation; they received support from their local randomised cluster and consented to follow-up data collection only. Carers were told that the service was being evaluated but not told about the randomised clusters.

Interventions and procedures

The intervention is a person-centred, structured process of assessment and support that is practitioner facilitated, but carer led. It enables carers to identify and prioritise their unmet needs during routine support contacts by staff; and then collaboratively put in place tailored support to meet identified needs. The intervention includes: a needs assessment tool; an action plan; and a multistage person-centred framework for introducing and using them both. The intervention is delivered typically at home visits that also include stroke survivors being supported by the same staff member. Staff in all clusters were trained in the study processes but only those in intervention clusters were trained to implement this individualised approach, using instructional videos, role-play and workbooks. Implementation does not include change to local, external support services available to carers—although staff were encouraged to create service directories, in case signposting or referral was required. The intended intervention is illustrated in [figure 1](#), summarised in [table 1](#) which highlights differences to usual care and described in detail in online supplemental table S1, adapted Template for Intervention Description and Replication checklist.²⁰

We compared the intervention to usual care within clusters (also summarised in [table 1](#) and described in online supplemental table S1). Although the service delivery organisation had well-defined practices for supporting stroke survivors, support for carers was typically offered but variable across services.

Study entry data included demographic and clinical characteristics of carers and their cared-for stroke

survivors, along with EQ-5D-5L.²¹ These were collected through carer self-report postal questionnaires at the same time as consent. As support (intervention or control) was implemented at a cluster level and designed to begin at the first point of contact with a carer, study entry data could not be considered truly 'baseline' as it was collected after support had been initiated, although data such as age, gender and date of stroke could be assumed to be constant. Initial and follow-up outcomes were sought by carer self-report postal questionnaire 3 and 6 months after support was initiated. In addition, service delivery records for all consented carers were extracted by the service provider at the end-of-study data collection.

Outcomes

Primary outcome was the strain subscale from the FACQ⁹ 3 months after the start of intervention. Three-month and 6-month outcomes postal questionnaire packs were identical in content. Carers were provided with the option to complete them over the telephone with support from a researcher. Packs included:

- ▶ The FACQ⁹ with subscales for strain, distress and positive appraisals of the impact of caring. Each item was scored from 1 to 5 and each subscale produced a mean score out of 5, with a score of 3 as neutral, and higher scores indicating a greater amount of the variable being measured.
- ▶ The Pound Carer Satisfaction with Stroke Services Scale,²² with higher scores indicating more satisfaction with services (composite score maximum of 44; standalone 'smiley faces' overall score maximum of 7).
- ▶ The Hospital Anxiety and Depression Scale²³ for carer anxiety and depression, with higher scores indicating higher mood disturbance and clinical cut-offs of: non-cases (0–7); mild (8–10); moderate (11–14); severe (15–21).
- ▶ An adapted version of the Service Receipt Inventory¹¹ to collect information on carers' use of NHS and social care services and the EQ-5D-5L²¹ as the measure of health benefit used in the economic evaluation.

Routinely collected service delivery records for consented carers (described in Interventions and Procedures) included: the dates, types and duration of direct and non-direct support activities provided; standardised entries from staff pertaining to needs identified and actions taken during support contacts. Needs and action categories were pre-existing within the service provider records management system and not altered for the purpose of the trial.

We collected data on how often staff used the intervention's needs assessment tool and action plan but primarily evaluated implementation using qualitative methods in our separately reported process evaluation.¹⁷

No serious adverse events (SAEs) were expected to be related to the intervention. All known AEs were typically collected via outcomes postal packs or during routine

study follow-up calls with participants or cluster staff. SAEs were reported if they were deemed related and unexpected. Protocol deviations were recorded, for example, return of 3-month outcome measures more than 6 weeks late.

Statistical and economic analysis

A full Statistical Analysis Plan was published with the study protocol.¹⁸ We explored a range of projected sample sizes in our protocol. A minimum of 400 carers recruited from 32 clusters (200 per trial arm) would provide 80% power to detect standardised effect sizes on the primary outcome of 0.31 or more (FACQ Strain mean score), assuming an intraclass correlation coefficient (ICC) of 0.01 with a 20% loss to follow-up, at the 5% significance level. Power was calculated using the Stata *clsampsi* function.²⁴ We did not expect the cluster ICC to be >0.05.²⁵

The primary analysis was intention to treat (ITT), comparing intervention and control at 3 months using a multilevel regression model with adjustment for clustering and using the following fixed individual level covariates: time post-stroke; age of carer; health of carer at study entry; stroke severity (as rated by carer); and the following cluster level covariates: size of service and experience of staff delivering support. Missing covariate data were imputed using multiple imputation via the 'mi impute' function in Stata. Sequential imputation using chained equations was used to create 10 datasets. At least 6 of the 8 items on the primary outcome (FACQ Strain subscale at 3 months) had to be completed for inclusion in primary analysis. Similar analysis was used for all numerical secondary outcome measures. Sensitivity analyses were prespecified in the Statistical Analysis Plan to explore any potential bias and examine the robustness of findings.

An analysis plan for the economic evaluation was also published as part of the study protocol. The economic evaluation compared the intervention with usual care over the 6-month follow-up period using an ITT approach and from the NHS and social care perspective. The measure of health benefit was utility, derived from EQ-5D-5L at each assessment using the crosswalk methods as currently recommended by the National Institute for Health and Care Excellence.²⁶ Quality-adjusted life years (QALYs) were calculated from these utility values using an area under the curve approach. The costs for the economic evaluation include the costs associated with NHS and social care resources used by carers during the study and the direct costs associated with delivering the intervention/control. The intervention-related costs included training for staff and time spent providing support (extracted from service delivery records). Further details of the economic methods are reported in online supplemental material.

Regression models, based on multiple imputed datasets, were used to estimate net costs (generalised linear model with gamma family and log link) and QALYs

(linear model) for the intervention arm compared with the control arm. Models allowed for clustering by adjusting for the same cluster-level covariates as the clinical-effectiveness analysis (see above) and the models were specified so that the CIs allowed for intragroup correlation. Net costs were divided by net QALYs to calculate an Incremental Cost-Effectiveness Ratio (ICER). The net costs and QALYs were bootstrapped 2000 times to estimate robust 95% CIs and plotted on a cost-effectiveness plane. Prespecified sensitivity analyses were conducted, including complete case analyses.

Role of the funding source

The NIHR CLAHRC had no role in study design, data collection, data analysis, data interpretation or writing of the paper. Stroke Association partnered with NIHR in funding this study and was the specialist stroke service provider in OSCARSS. They did provide some data (eg, service delivery records) and contributed to discussions about data interpretation and dissemination of findings. The corresponding author had full access to all the

study data and had final responsibility for the decision to submit for publication.

RESULTS

In September 2016 we randomised 36 clusters (18 intervention; 18 control). Three control and one intervention cluster withdrew soon after due to decommissioning or all staff long-term sickness (see figure 2) so 32 clusters were trained in January 2017. Three replacement clusters were recruited, randomised and trained between February and April 2017 (one intervention; two control). This gave a total of 35 recruiting clusters (18 intervention; 17 control). Cluster and staff baseline characteristics are included in online supplemental table S2.

Between January 2017 and July 2018, 628 eligible carers (334 intervention; 294 control) were referred for potential participation across 35 participating clusters (18 intervention; 17 control) in England and Northern Ireland. Of those eligible, 414 (66%) carers consented (208 intervention; 206 control) and were followed up between March

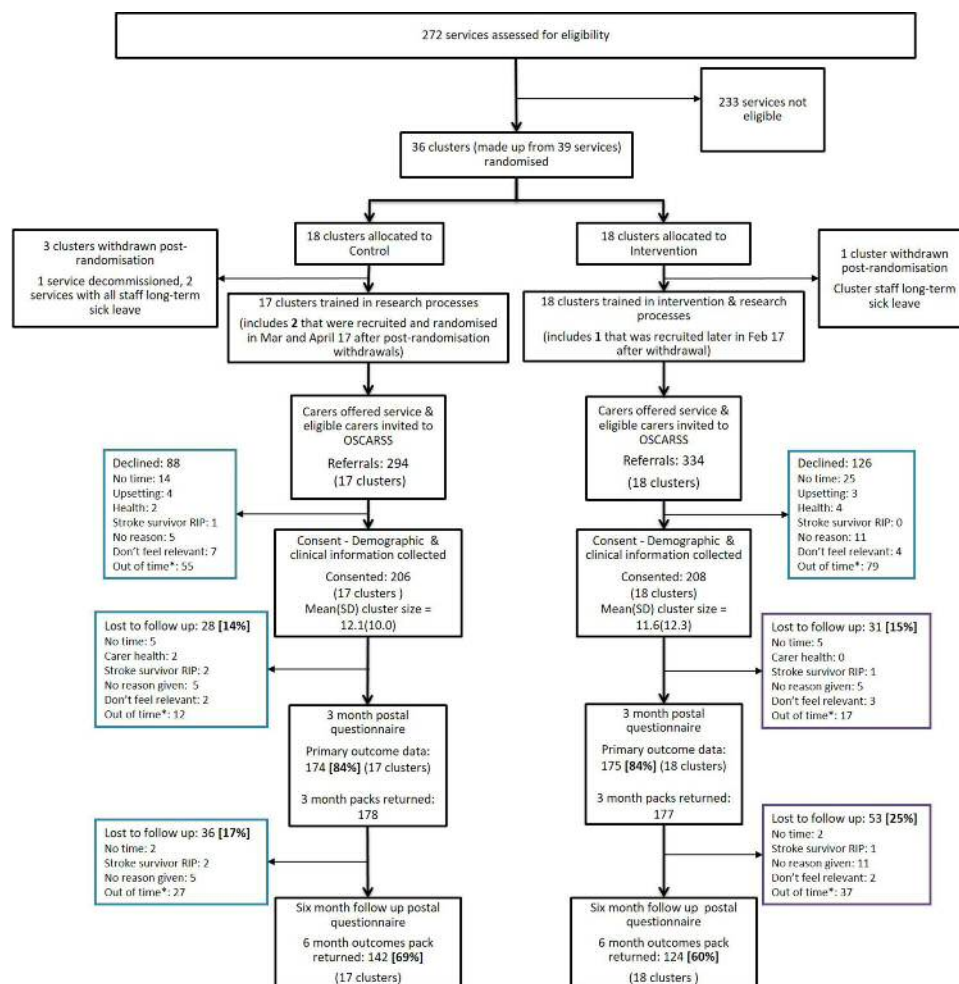


Figure 2 CONSORT diagram showing cluster recruitment and patient flow. All numbers correspond to number of carers unless otherwise stated. All percentages are out of number of consented carers. *Postal packs had not been returned after 13 weeks/21 weeks/26 weeks for demographic/3-month/6-month data. CONSORT, Consolidated Standards of Reporting Trials; OSCARSS, Organising Support for Carers of Stroke Survivors.

2017 and December 2018. Participant flow is shown in [figure 2](#) and consented carer study entry characteristics are shown in [table 2](#).

Of the 414 consented participants, 319 (77%) were women and 315 (76%) were partners/spouses of the stroke survivor they cared for and 399 (97%) were ethnically white. The mean age of carers was 62 years old when they joined the study and the median time from the stroke event to support being initiated was 2.3 months across the whole sample (IQR=1.1–2.3). All measured variables related to consented carers were well balanced across intervention and control groups, including the level of independence of the cared-for stroke survivor, as perceived by carers.

Primary outcomes were available for 175 (84%) of consented carers in the intervention group and 174 (84%) in the control group. Follow-up (secondary) outcomes were available for 124 (60%) of consented carers in the intervention group and 142 (69%) in the control group.

Primary analysis of all outcomes is shown in [table 3](#). Clustering for the primary outcome was low (ICC=0.02) and negligible after adjustment for covariates. For our primary outcome measure we found the mean (SD) FACQ carer strain at 3 months to be 3.11 (0.87) in the control group compared with 3.03 (0.90) in the intervention group, adjusted mean difference –0.04 (95% CI –0.20 to 0.13). Note that this CI excludes the minimal important difference of 0.31 used in our sample size calculation and therefore the data are not consistent with a clinically relevant difference between intervention and control groups. Similarly when we looked at the longer term FACQ carer strain at 6 months we observed a mean control measure of 3.10 (0.88) compared with 3.07 (0.87), adjusted mean difference –0.04 (95% CI –0.22 to 0.14). All other secondary outcome measures had small differences and tight CIs (see [table 3](#)) and therefore are not consistent with meaningful differences between control and intervention. Both unadjusted and adjusted estimates of intervention effect were similar, providing no evidence of any confounding due to demographic or clinical variables. These findings were consistent across all sensitivity analyses including: excluding delayed responders; removing carer dyads; imputing missing outcome data; and combining 3-month and 6-month data, suggesting that the results are robust to assumptions made in the analysis.

The clinical interpretation of selected findings was that for the primary outcome, carer strain, both groups reported an average of around 3 out of 5 that is, a neutral level. For secondary outcomes, average levels of anxiety and depression were around 8 and 6/7 out of 21 (mild and non-case, respectively). Both groups tended to ‘agree’ with the positive appraisal of the impact of caregiving that is, average scores 4 out of 5. Satisfaction ratings for both groups were towards the higher end of the composite scale, an average of around 30 out of 44.

For the economic evaluation, there was a high proportion of missing data but economic analysis was still

Table 2 Carer study entry characteristics

	Control	Intervention
	N=206	N=208
Sex, n (%)		
Male	42 (20.4)	51 (24.5)
Female	164 (79.6)	155 (74.5)
Missing data	–	2 (1)
Age, mean (range)		
	62.5 (24–86)	62.3 (21–88)
Relationship with stroke survivor, n (%)		
Husband/wife or partner	160 (77.7)	155 (74.5)
Parent	2 (1.0)	6 (2.9)
Son/daughter	39 (18.9)	41 (19.7)
Other	5 (2.5)	5 (2.5)
Missing data	–	1 (0.5)
Lives relative to stroke survivor, n (%)		
In the same household	179 (86.9)	172 (82.7)
Within walking distance	8 (3.9)	12 (5.8)
Within 30 min drive/public transport	16 (7.8)	16 (7.7)
More than 30 min drive/public transport	3 (1.5)	8 (3.8)
Marital status, n (%)		
Single	18 (9)	9 (4)
Married/living as married	177 (85)	178 (87)
Other	13 (6)	18 (8)
Missing data	1 (0)	–
Ethnicity, n (%)		
White	200 (97.1)	199 (96.7)
Mixed/multiple ethnic groups	–	4 (1.9)
Asian/Asian British	6 (2.9)	5 (2.4)
Employment status, n (%)		
Employed full-time	30 (14.6)	25 (12.0)
Employed part-time	23 (11.2)	25 (12.0)
Self-employed	13 (6.3)	9 (4.3)
Retired	102 (49.5)	111 (53.4)
Unemployed	11 (5.3)	12 (5.8)
Full-time education	–	1 (0.5)
Other, including homemaker	27 (13.1)	25 (12)
Highest level of education, n (%)		
None	47 (22.8)	49 (23.6)
Examinations at 16	72 (35.0)	75 (36.1)
A/AS level or equivalent	41 (19.9)	29 (13.9)
University	42 (20.4)	50 (24.0)

Continued

Table 2 Continued

	Control N=206	Intervention N=208
Other	1 (0.5)	2 (1.0)
Missing data	3 (1.5)	3 (1.4)
Carer has long-term health condition, n (%)		
Yes	124 (60.2)	130 (62.5)
No	82 (39.8)	78 (37.5)
Carer provided care to stroke survivor prior to stroke, n (%)		
Yes	81 (39.3)	79 (38.0)
No	124 (60.2)	122 (58.7)
Cared-for stroke survivor characteristics (as reported by carer)		
Months post-stroke (at date seen)		
Mean (SD)	5.93 (15.47)	6.46 (16.38)
Median (IQR)	2.2 (1.1–4.6)	2.37 (1.2–4.8)
Missing data, n (%)	3 (1.5)	8 (3.8)
Independence*		
Mean (SD)	10.99 (3.67)	11.14 (3.69)
Median (IQR)	11 (8–14)	11 (8–14)

*Mean score for carer perceived independence calculated over 6 domains: personal care, toilet, cooking, walking, transport and finances/legal issues. Each domain scored 1–3 (total max score=18) with low scores equating to greater independence.

feasible. We found similar neutral findings between groups in terms of health benefits (see [table 4](#)). Resource use is summarised in online supplemental tables S3–S8. Costs associated with the intervention were slightly higher (around £40 per person) than the control, primarily due to:

- ▶ Additional staff training required for the intervention, calculated at £15 per consented carer supported in intervention-allocated clusters.
- ▶ Additional support provided to consented carers in intervention-allocated clusters, according to extracted service delivery records. Carers in intervention-allocated versus control-allocated clusters had 15 vs 12 support activities recorded, on average, totalling 4.7 hours vs 4.2 hours, respectively.
- ▶ Carers in intervention-allocated clusters self-reported accessing more primary care services, specifically general practice nurses.

These slightly higher costs without measurable health benefits over usual care suggest that the intervention as delivered is unlikely to be cost-effective (see [table 5](#)). This remained the case in all sensitivity analyses. [Figure 3](#) shows the cost-effectiveness plane for the primary analysis; the clustering around the vertical axis demonstrates that we can be relatively certain there is no additional health benefit from the intervention compared with the control group.

No SAEs were reported that were judged to be related to the research. There were 12 SAEs in total (seven intervention; five control). Ten involved hospitalisation and two related to Accident and Emergency visits with possible long-term incapacity.

Service delivery records indicate that more carers received an individual case record in the intervention arm (92/208, 44%) than control arm (65/206, 32%); other carers had service delivery data captured alongside a stroke survivor record. In addition, intervention arm carers versus control arm carers had more needs reported (146 vs 80) and more actions agreed (278 vs 148), according to service delivery records.

Indicative findings from the quantitative data on the implementation of the intervention suggest the intervention was not implemented as intended. Overall, of the 334 eligible carers referred to the study from intervention-allocated clusters, the CSNAT-Stroke needs assessment tool and action plan were recorded as used in 278 (83%) and 121 (36%) cases, respectively. Similarly, for the 208/334 carers who went on to join the study from intervention-allocated clusters, they were used in 172 (83%) and 66 (32%) cases, respectively.

DISCUSSION

In terms of clinical effectiveness and cost-effectiveness OSCARSS' findings were conclusive. We found no meaningful difference in the level of self-reported caregiver strain between those allocated to an adapted support intervention or to usual care. Findings were robust and consistent across all outcomes, time points and sensitivity analyses. The economic evaluation demonstrated neutral findings on health benefits and slightly increased costs making the intervention unlikely to be cost-effective compared with usual care. There are several possible explanations for our neutral finding explored in detail below. In brief, carers in both groups received support from the same national service provider organisation, and at the primary outcome time point both groups had a level of strain categorised as neutral on average. Carers seen by intervention-allocated clusters received slightly more support and accessed more primary care services than carers in the control group. However, the intervention was not fully delivered as intended.

Comparison with other studies

A review of multifaceted support interventions for stroke survivors and carers found no evidence of effectiveness for carers' subjective health status nor mental health (15 interventions, 1775 carers).¹⁴ A review of non-pharmacological interventions for carers of stroke survivors also found no strong evidence to inform best practice for supporting carers (8 studies, 1007 carers).¹³ Recent important randomised trials of structured training for carers to provide care¹¹ or deliver rehabilitation¹² show the feasibility of carer trials but an absence of evidence of effectiveness.

**Table 3** Primary analysis of all outcomes

	Control	Intervention	Difference (95% CI)
	Mean (SD)	Mean (SD)	Adjusted for clustering and demographic variables
Primary outcome: FACQ carer strain at 3 months	N=174 3.11 (0.87)	N=175 3.03 (0.90)	-0.04 (-0.20 to 0.13)
Secondary outcomes collected at 3 months after support initiated:			
FACQ carer distress	N=173 2.88 (0.83)	N=176 2.91 (0.85)	0.04 (-0.13 to 0.21)
FACQ positive caregiving appraisal	N=175 3.99 (0.61)	N=176 4.05 (0.54)	0.05 (-0.06 to 0.17)
Pound Satisfaction with stroke services (composite)	N=177 31.14 (8.85)	N=171 30.51 (10.36)	-1.06 (-3.35 to 1.23)
Pound overall Satisfaction with stroke services (smiley faces)	N=174 5.10 (1.51)	N=167 5.10 (1.49)	0.00 (-0.30 to 0.31)
HADS anxiety	N=174 8.34 (4.51)	N=172 8.20 (4.73)	0.04 (-0.89 to 0.97)
HADS depression	N=174 6.30 (4.17)	N=172 6.12 (4.07)	-0.06 (-0.86 to 0.73)
Follow-up outcomes collected at 6 months after support initiated:			
FACQ carer strain	N=140 3.10 (0.88)	N=121 3.07 (0.87)	-0.04 (-0.22 to 0.14)
FACQ carer distress	N=140 2.93 (0.84)	N=121 2.92 (0.84)	-0.03 (-0.23 to 0.16)
FACQ positive caregiving appraisal	N=140 3.91 (0.64)	N=121 4.04 (0.54)	0.12 (-0.02 to 0.26)
Pound Satisfaction with stroke services (composite)	N=136 32.12 (5.88)	N=121 30.58 (9.81)	-1.48 (-3.40 to 0.44)
Pound overall Satisfaction with stroke services (smiley faces)	N=138 5.17 (1.51)	N=120 4.99 (1.54)	-0.21 (-0.61 to 0.20)
HADS anxiety	N=141 8.90 (4.66)	N=123 8.95 (5.10)	0.13 (-0.98 to 1.23)
HADS depression	N=141 7.06 (4.56)	N=123 6.65 (4.06)	-0.43 (-1.36 to 0.51)

FACQ, Family Appraisal of Caregiving Questionnaire; HADS, Hospital Anxiety and Depression Scale.

Prior to OSCARSS there was no robust RCT evidence of the CSNAT, or any other approach, to guide the support of carers of stroke survivors. A non-randomised study of CSNAT with a non-stroke population concluded that the CSNAT was associated with small to moderate reductions in carer strain compared with pre-intervention.^{6 8} Several UK studies by the CSNAT team showed similar outcomes and good acceptability, but also reported implementation challenges similar to those found in OSCARSS^{5 7} and discussed in our sister process evaluation paper.¹⁷

Strengths and limitations, with consideration of clinical implications

To understand the clinical implications of these findings, we consider the study's strengths and limitations, and explore possible explanatory factors: the choice of comparator; intervention delivered; the timing and choice of outcomes and characteristics of the sample.

Features of the study design and conduct ensured good internal validity. For example, clusters were recruited prior to stratified randomisation and carer research participants—who completed self-reported outcome

Table 4 EQ-5D utility values at each time point and QALYs for whole follow-up, by treatment arm

	Control	Intervention
	Mean (95% CI)	
Study entry utility	0.78 (0.75 to 0.81) n=204	0.76 (0.74 to 0.79) n=199
3-month utility	0.73 (0.71 to 0.76) n=177	0.73 (0.70 to 0.76) n=165
6-month utility	0.72 (0.69 to 0.75) n=136	0.73 (0.69 to 0.76) n=118
QALYs (over 6 months)	0.37 (0.36 to 0.38) n=135	0.38 (0.36 to 0.39) n=103
Net QALYs*	0.009 (−0.016 to 0.033) n=238	
Adjusted net QALYs†	0.004 (−0.018 to 0.026) n=227	

*Unadjusted but allowing for intracluster correlation in SEs.

†Net QALYs calculated using linear regression model adjusted for age, time since stroke, stroke severity, whether or not the carer had any long-term health conditions, cluster size and years of experience of the cluster staff. QALYs, quality-adjusted life years.

measures—were unaware of allocation. Cluster trials risk imbalance across trial groups²⁷ but in OSCARSS, steps were taken to minimise this and all measured variables related to consented carers appeared well-balanced across arms. The intervention was implemented at cluster level and began the moment a carer came into

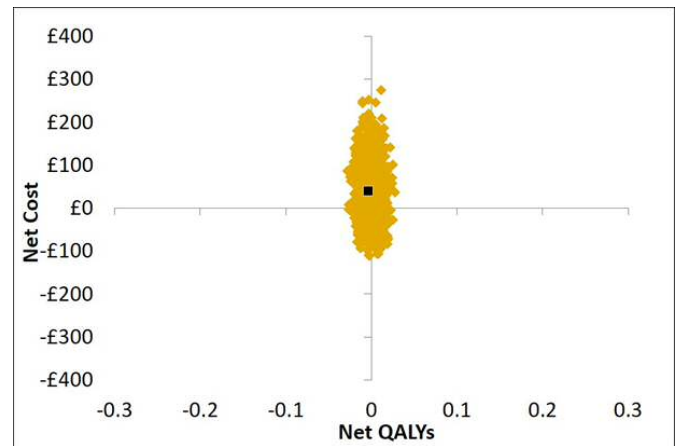


Figure 3 Cost-effectiveness plane for primary analysis. The cost-effectiveness plane shows the ICER (large square) and 2000 bootstrapped estimates of net costs and QALYs. The narrow, even, horizontal spread of the points indicates low uncertainty regarding the indifferent health benefit. The broader vertical spread of the points shows that there is more uncertainty around the costs. ICER, incremental cost-effectiveness ratio; QALYs, quality-adjusted life years.

contact with the service provider so it was not possible to explore change from baseline in individual outcomes, however the randomised design, coupled with balanced cluster and carer characteristics, helps overcome this. OSCARSS achieved its target, powered sample size with minimal missing clinical data and low attrition (16% in both groups) at the primary outcome time point. We have confidence in our findings which were consistent across

Table 5 Results of primary and sensitivity economic analyses comparing CSNAT intervention with usual care

	Net costs (95% CI)	Net QALYs (95% CI)	ICER (£/QALY)
Primary analysis			
Multiple imputed datasets (n=410)*	£39.05 (−69.61 to 147.71)	−0.004 (−0.020 to 0.012)	Intervention is dominated
Sensitivity analyses			
Complete cases (n=131)	£41.24 (−29.01 to 111.49)	−0.0001 (−0.026 to 0.026)	Intervention is dominated
Per-protocol analysis† (n=374)	£42.55 (−71.77 to 156.88)	−0.0002 (−0.016 to 0.016)	Intervention is dominated
Exclude training and intervention costs (n=410)*	£23.33 (−98.21 to 144.87)	−0.004 (−0.020 to 0.012)	Intervention is dominated
Alternative outcome measure			
	Net costs (95% CI)	Net change (95% CI)	ICER: (£/1 point improvement)
FACQ strain, complete cases (n=139)	£57.32 (−15.77 to 130.41)	−0.02 (ie, lower score in intervention group) (−0.30 to 0.26)	Intervention is dominated

All analyses adjusted for covariates: carer's age, time since stroke, stroke severity, whether or not carer has long-term health conditions, length of experience of cluster staff, size of cluster and cluster ID.

CIs for all analyses calculated following bootstrapping: 2000 times for imputed datasets, 10 000 times for complete case datasets.

*Four participants with no baseline EQ-5D data were excluded from the imputation, leaving 410 participants.

†Thirty-six participants in the imputed dataset excluded who violated protocol conditions (multiple carers per stroke survivor or questionnaires returned late).

CSNAT, Carer Support Needs Assessment Tool; FACQ, Family Appraisal of Caregiving Questionnaire; ICER, incremental cost-effectiveness ratio; QALYs, quality-adjusted life years.



all sensitivity analyses including for protocol deviations such as the late return of postal questionnaires.

The demographic profile of the sample was as expected for carers of stroke survivors and in keeping with other trials.¹¹ However, as is so often the case in UK-based stroke trials, the sample lacked ethnic diversity (<3% non-white group). This does not reflect the diversity in the UK general population. Stroke trials need strategies to achieve equity of access, given that a large portion of UK stroke admissions are from Black, Asian and minority ethnic communities.²⁸ We aimed to recruit the primary caregiver but did not collect additional data on whether they were caring alone or with support. All other measured carer variables were balanced across randomised groups.

Neutral findings must consider the context that carers in both groups received support, from the same stroke specialist provider organisation, and reported high satisfaction with stroke services on average. This and the outcomes achieved suggest it is plausible that both methods of support delivered in OSCARSS were beneficial to carers.

We collaborated closely with our service provider to pragmatically tailor the intervention for implementation, which improved buy-in by the organisation and cluster staff. However, our data show that the intervention's assessment tool and action plan were underused. Implementation was explored in greater depth in the embedded process evaluation and is consistent with these quantitative indicators; namely, that the intervention as intended was not fully implemented.¹⁷

We have no data beyond 6 months after support had been initiated, and while our inclusion criteria aimed to recruit carers at varying stages, our sample was predominantly early post-stroke. Previous stroke research suggests caregivers may take months to adjust to their role as caregivers, become aware of and prioritise their own needs.²⁹ The OSCARSS process evaluation and opinions of members of our study-specific carer advisory research group endorse this and suggest that, while informal caregivers need support early after stroke, they may struggle to participate fully in a 'carer-led' intervention that encourages self-management, such as the CSNAT intervention, which could have contributed to the implementation issues noted above. In addition, our relatively short follow-up period of 6 months may have been too early to detect any impact of carers in the intervention group receiving more support and accessing more primary healthcare services, as observed in our economic evaluation. While our choice of primary outcome was informed by past research using the CSNAT intervention^{6,8} and the preferences of our service user group of stroke carers, our measure may not have been adequate to detect a difference in our population of stroke carers.

CONCLUSIONS

In summary, OSCARSS found that the CSNAT-Stroke intervention was not measurably clinically effective or

cost-effective compared with usual care from a stroke specialist provider organisation, although we have substantial evidence that the intervention was not fully implemented in this pragmatic trial. OSCARSS demonstrated that methodologically rigorous research evaluations for carers of stroke survivors can be successfully delivered by voluntary sector organisations. However, the challenges of fully implementing person-centred care in research and service development need to be addressed through enhanced and ongoing staff training as well as organisational mechanisms to support and champion new approaches becoming embedded into practice. There remains a high priority for research to determine how best to support carers of stroke survivors.

Author affiliations

¹Division of Neuroscience and Experimental Psychology, The University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK

²National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Greater Manchester (NIHR CLAHRC GM), Manchester, UK

³Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, The University of Manchester, Manchester, UK

⁴Manchester Centre for Health Economics, Division of Population Health, Health Services Research & Primary Care, School of Health Sciences, The University of Manchester, Manchester, UK

⁵Alliance Manchester Business School, The University of Manchester, Manchester, UK

⁶Centre for Reviews and Dissemination, University of York, York, UK

⁷Division of Nursing Midwifery and Social Work, School of Health Sciences, The University of Manchester, Manchester Academic Health Sciences Centre (MAHSC), Manchester, UK

⁸Centre for Family Research, University of Cambridge, Cambridge, UK

Twitter Emma Patchwood @DrPatchwood, Sarah Darley @sardarl, Gunn Grande @gunn_grande, Gail Ewing @gaillewing_cfr and Audrey Bowen @audreybowenprof

Acknowledgements The authors would like to thank all of the carers and staff members who participated and contributed to this study. We would also like to thank and acknowledge members of the supportive NIHR CLAHRC GM OSCARSS Research Team who helped make the study possible. This includes members of the OSCARSS carer advisory Research User Group (Kelly Burke, Christine Halford, Natalie Halford, Geoff Heathcote, Kath Purcell and Ben Wright); members from the trial management group (Caroline O'Donnell (data analyst), Alison Littlewood and Katy Rothwell (programme managers), Sam Wilkinson, Amy Woodhouse and Rose Crees (administrative assistants), Aneela Macavoy and Zoe Ashton (research facilitators)); and members of the Trial Steering Committee (Chris Sutton (chair and statistician), David Clarke (process evaluation steer), Rachael Hunter (health economics steer), Jordi Morell (clinical steer) and Nigel Bamford (lay member)).

Contributors EP and AB are co-chief investigators of OSCARSS and conceived the study. EP wrote the first draft of the paper and all authors substantially contributed to revisions. KW-N project managed OSCARSS. SAR and EB performed statistical analysis. EC performed economic analysis. SK and SD led the process evaluation in OSCARSS and contributed to interpretation of the implementation data presented in this paper. GG and GE developed the original CSNAT intervention and contributed substantially to descriptions of the adapted CSNAT-Stroke intervention in this paper.

Funding This study was funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Greater Manchester (NIHR CLAHRC GM), grant number (N/A), partnered with Stroke Association Funding, grant number (N/A).

Competing interests AB, GG, SAR, EB and GE held grants with NIHR during the course of the OSCARSS Study. AB and EP additionally hold grants with Stroke Association outside of this work. There are no other relationships or activities that could appear to have influenced the submitted work.

Patient consent for publication Not required.

Ethics approval Ethics approvals were obtained from Lancaster Research Ethics Committee (ref: 16/NW/0657).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Requests for data and statistical code should be made to the corresponding author and will be considered by members of the original trial management group, including the co-chief investigators, who will release data on a case-by-case basis. Data will be shared following the principles for sharing patient-level data as described by Smith *et al* (2015). The data will not contain any direct identifiers, we will minimise indirect identifiers and remove free text data, to minimise the risk of identification.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) 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, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Emma Patchwood <http://orcid.org/0000-0002-4198-5761>

Elizabeth Camacho <http://orcid.org/0000-0001-9574-7710>

Sarah Darley <http://orcid.org/0000-0001-5420-6774>

Gunn Grande <http://orcid.org/0000-0003-2200-1680>

Gail Ewing <http://orcid.org/0000-0001-9547-7247>

Audrey Bowen <http://orcid.org/0000-0003-4075-1215>

REFERENCES

- Anderson R. The unremitting burden on carers. *BMJ* 1987;294:73–4.
- Carers UK. State of caring report, 2019. Available: www.carersuk.org/stateofcaring-report
- Department of Health. Care act, 2014. Available: <http://www.legislation.gov.uk/ukpga/2014/23/contents/enacted/data.htm>
- Ewing G, Grande G, National Association for Hospice at Home. Development of a carer support needs assessment tool (CSNAT) for end-of-life care practice at home: a qualitative study. *Palliat Med* 2013;27:244–56.
- Aoun S, Deas K, Toye C, *et al*. Supporting family caregivers to identify their own needs in end-of-life care: qualitative findings from a stepped wedge cluster trial. *Palliat Med* 2015;29:508–17.
- Aoun SM, Grande G, Howling D, *et al*. The impact of the carer support needs assessment tool (CSNAT) in community palliative care using a stepped wedge cluster trial. *PLoS One* 2015;10:e0123012.
- Ewing G, Austin L, Grande G. The role of the carer support needs assessment tool in palliative home care: a qualitative study of practitioners' perspectives of its impact and mechanisms of action. *Palliat Med* 2016;30:392–400.
- Grande GE, Austin L, Ewing G. Assessing the impact of a carer support needs assessment tool (CSNAT) intervention in palliative home care: a stepped wedge cluster trial. *BMJ Support Palliat Care* 2017;7:326–34.
- Cooper B, Kinsella GJ, Picton C. Development and initial validation of a family appraisal of caregiving questionnaire for palliative care. *Psychoncology* 2006;15:613–22.
- Adamson J, Beswick A, Ebrahim S. Is stroke the most common cause of disability? *J Stroke Cerebrovasc Dis* 2004;13:171–7.
- Forster A, Dickerson J, Young J, *et al*. A structured training programme for caregivers of inpatients after stroke (TRACS): a cluster randomised controlled trial and cost-effectiveness analysis. *Lancet* 2013;382:2069–76.
- Lindley RI, Anderson CS, Billot L, *et al*. Family-led rehabilitation after stroke in India (attend): a randomised controlled trial. *Lancet* 2017;390:588–99.
- Legg LA, Quinn TJ, Mahmood F, *et al*. Non-pharmacological interventions for caregivers of stroke survivors. *Cochrane Database Syst Rev* 2011;10:CD008179.
- Ellis G, Mant J, Langhorne P, *et al*. Stroke liaison workers for stroke patients and carers: an individual patient data meta-analysis. *Cochrane Database Syst Rev* 2010;17:CD005066.
- Cameron JI, Naglie G, Gignac MAM, *et al*. Randomized clinical trial of the timing of right stroke family support program: research protocol. *BMC Health Serv Res* 2014;14:18.
- Mitchell C, Burke K, Halford N, *et al*. Value and learning from carer involvement in a cluster randomised controlled trial and process evaluation - Organising Support for Carers of Stroke Survivors (OSCARSS). *Res Involv Engagem* 2020;6.
- Darley S, Knowles S, Woodward- Nutt K, *et al*. Challenges implementing a carer support intervention within a national stroke organisation: findings from the process evaluation of the OSCARSS trial. *BMJ Open* 2020.
- Patchwood E, Rothwell K, Rhodes S, *et al*. Organising support for carers of stroke survivors (OSCARSS): study protocol for a cluster randomised controlled trial, including health economic analysis. *Trials* 2019;20:19.
- Staniszewska S, Brett J, Simera I, *et al*. GRIPP2 reporting checklists: tools to improve reporting of patient and public involvement in research. *Res Involv Engagem* 2017;3:13.
- Hoffmann TC, Glasziou PP, Boutron I, *et al*. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687.
- Herdman M, Gudex C, Lloyd A, *et al*. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011;20:1727–36.
- Pound P, Gompertz P, Ebrahim S. Development and results of a questionnaire to measure carer satisfaction after stroke. *J Epidemiol Community Health* 1993;47:500–5.
- Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67:361–70.
- Batistatou E, Roberts C, Roberts S. Sample size and power calculations for trials and quasi-experimental studies with clustering. *Stata J* 2014;14:159–75.
- Campbell MK, Thomson S, Ramsay CR, *et al*. Sample size calculator for cluster randomized trials. *Comput Biol Med* 2004;34:113–25.
- van Hout B, Janssen MF, Feng Y-S, *et al*. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 2012;15:708–15.
- Farrin A, Russell I, Torgerson D, *et al*. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK beam) feasibility study. *Clin Trials* 2005;2:119–24.
- Stroke Association. State of the nation: stroke statistics. Available: https://www.stroke.org.uk/sites/default/files/state_of_the_nation_2017_final_1.pdf 2017
- Cameron JI, Gignac MAM. "Timing It Right": a conceptual framework for addressing the support needs of family caregivers to stroke survivors from the hospital to the home. *Patient Educ Couns* 2008;70:305–14.

6.3 PAPER 3

Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. Cochrane Database of Systematic Reviews. 2019(8). <https://doi.org/10.1002/14651858.CD012248.pub2>

(Characteristics of studies and some appendices not included).



Cochrane
Library

Cochrane Database of Systematic Reviews

Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection (Review)

Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF

Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF.
Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection.
Cochrane Database of Systematic Reviews 2019, Issue 8. Art. No.: CD012248.
DOI: [10.1002/14651858.CD012248.pub2](https://doi.org/10.1002/14651858.CD012248.pub2).

www.cochranelibrary.com

TABLE OF CONTENTS

HEADER	1
ABSTRACT	1
PLAIN LANGUAGE SUMMARY	2
SUMMARY OF FINDINGS	4
BACKGROUND	6
OBJECTIVES	7
METHODS	7
RESULTS	10
Figure 1.	11
Figure 2.	14
Figure 3.	15
DISCUSSION	18
AUTHORS' CONCLUSIONS	19
ACKNOWLEDGEMENTS	19
REFERENCES	20
CHARACTERISTICS OF STUDIES	24
DATA AND ANALYSES	39
Analysis 1.1. Comparison 1 Chlorhexidine bathing versus soap-and-water bathing, Outcome 1 Hospital-acquired infection. ...	40
Analysis 1.2. Comparison 1 Chlorhexidine bathing versus soap-and-water bathing, Outcome 2 Mortality using adjusted data. ..	41
ADDITIONAL TABLES	41
APPENDICES	42
HISTORY	51
CONTRIBUTIONS OF AUTHORS	51
DECLARATIONS OF INTEREST	51
SOURCES OF SUPPORT	51
DIFFERENCES BETWEEN PROTOCOL AND REVIEW	52
INDEX TERMS	52

[Intervention Review]

Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection

Sharon R Lewis¹, Oliver J Schofield-Robinson¹, Sarah Rhodes², Andrew F Smith³

¹Lancaster Patient Safety Research Unit, Royal Lancaster Infirmary, Lancaster, UK. ²Division of Population Health, Health Services Research & Primary Care, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ³Department of Anaesthesia, Royal Lancaster Infirmary, Lancaster, UK

Contact address: Sharon R Lewis, Lancaster Patient Safety Research Unit, Royal Lancaster Infirmary, Pointer Court 1, Ashton Road, Lancaster, LA1 4RP, UK. Sharon.Lewis@mbht.nhs.uk, sharonlewis@googlemail.com.

Editorial group: Cochrane Wounds Group.

Publication status and date: Edited (no change to conclusions), published in Issue 3, 2020.

Citation: Lewis SR, Schofield-Robinson OJ, Rhodes S, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. *Cochrane Database of Systematic Reviews* 2019, Issue 8. Art. No.: CD012248. DOI: [10.1002/14651858.CD012248.pub2](https://doi.org/10.1002/14651858.CD012248.pub2).

Copyright © 2020 The Cochrane Collaboration. Published by John Wiley & Sons, Ltd.

ABSTRACT

Background

Hospital-acquired infection is a frequent adverse event in patient care; it can lead to longer stays in the intensive care unit (ICU), additional medical complications, permanent disability or death. Whilst all hospital-based patients are susceptible to infections, prevalence is particularly high in the ICU, where people who are critically ill have suppressed immunity and are subject to increased invasive monitoring. People who are mechanically-ventilated are at infection risk due to tracheostomy and reintubation and use of multiple central venous catheters, where lines and tubes may act as vectors for the transmission of bacteria and may increase bloodstream infections and ventilator-associated pneumonia (VAP). Chlorhexidine is a low-cost product, widely used as a disinfectant and antiseptic, which may be used to bathe people who are critically ill with the aim of killing bacteria and reducing the spread of hospital-acquired infections.

Objectives

To assess the effects of chlorhexidine bathing on the number of hospital-acquired infections in people who are critically ill.

Search methods

In December 2018 we searched the Cochrane Wounds Specialised Register; the Cochrane Central Register of Controlled Trials (CENTRAL); Ovid MEDLINE; Ovid Embase and EBSCO CINAHL Plus. We also searched clinical trial registries for ongoing and unpublished studies, and checked reference lists of relevant included studies as well as reviews, meta-analyses and health technology reports to identify additional studies. There were no restrictions with respect to language, date of publication or study setting.

Selection criteria

We included randomised controlled trials (RCTs) that compared chlorhexidine bathing with soap-and-water bathing of patients in the ICU.

Data collection and analysis

Two review authors independently assessed study eligibility, extracted data and undertook risk of bias and GRADE assessment of the certainty of the evidence.

Main results

We included eight studies in this review. Four RCTs included a total of 1537 individually randomised participants, and four cluster-randomised cross-over studies included 23 randomised ICUs with 22,935 participants. We identified one study awaiting classification, for which we were unable to assess eligibility.

The studies compared bathing using 2% chlorhexidine-impregnated washcloths or dilute solutions of 4% chlorhexidine versus soap-and-water bathing or bathing with non-antimicrobial washcloths.

Eight studies reported data for participants who had a hospital-acquired infection during the ICU stay. We are uncertain whether using chlorhexidine for bathing of critically ill people reduces the rate of hospital-acquired infection, because the certainty of the evidence is very low (rate difference 1.70, 95% confidence interval (CI) 0.12 to 3.29; 21,924 participants). Six studies reported mortality (in hospital, in the ICU, and at 48 hours). We cannot be sure whether using chlorhexidine for bathing of critically-ill people reduces mortality, because the certainty of the evidence is very low (odds ratio 0.87, 95% CI 0.76 to 0.99; 15,798 participants). Six studies reported length of stay in the ICU. We noted that individual studies found no evidence of a difference in length of stay; we did not conduct meta-analysis because data were skewed. It is not clear whether using chlorhexidine for bathing of critically ill people reduced length of stay in the ICU, because the certainty of the evidence is very low. Seven studies reported skin reactions as an adverse event, and five of these reported skin reactions which were thought to be attributable to the bathing solution. Data in these studies were reported inconsistently and we were unable to conduct meta-analysis; we cannot tell whether using chlorhexidine for bathing of critically ill people reduced adverse events, because the certainty of the evidence is very low.

We used the GRADE approach to downgrade the certainty of the evidence of each outcome to very low. For all outcomes, we downgraded evidence because of study limitations (most studies had a high risk of performance bias, and we noted high risks of other bias in some studies). We downgraded evidence due to indirectness, because some participants in studies may have had hospital-acquired infections before recruitment. We noted that one small study had a large influence on the effect for hospital-acquired infections, and we assessed decisions made in analysis of some cluster-randomised cross-over studies on the effect for hospital-acquired infections and for mortality; we downgraded the evidence for these outcomes due to inconsistency. We also downgraded the evidence on length of stay in the ICU, because of imprecision. Data for adverse events were limited by few events and so we downgraded for imprecision.

Authors' conclusions

Due to the very low-certainty evidence available, it is not clear whether bathing with chlorhexidine reduces hospital-acquired infections, mortality, or length of stay in the ICU, or whether the use of chlorhexidine results in more skin reactions.

PLAIN LANGUAGE SUMMARY

Bathing critically ill patients with chlorhexidine to prevent hospital-acquired infections

What is the aim of this review?

The aim of this review was to find out whether people who are critically ill in hospital should be bathed with the antiseptic chlorhexidine, in order to prevent them from developing infections. Researchers from Cochrane collected and analysed all relevant studies to answer this question and found eight relevant randomised trials. Randomised trials are medical studies where people are chosen at random to receive different treatments. This study design provides the most reliable evidence on whether treatments have a relationship with desired or undesired health outcomes.

Key messages

This review assesses whether using chlorhexidine (instead of soap and water) to bathe patients in an intensive care unit (ICU), or a high-dependency or critical care unit reduces the number of hospital-acquired infections. The evidence available from the studies we analysed was very low quality, meaning that we cannot be certain whether bathing with chlorhexidine reduces the likelihood of critically-ill patients developing an infection, or dying. We are also uncertain whether bathing critically ill patients with chlorhexidine shortens the length of time people spend in hospital, or lowers their risk of developing skin reactions.

What was studied in the review?

People who are critically ill (in an ICU, or a high-dependency or critical care unit) often catch infections during their time in hospital. These infections can lead to longer hospital stays, additional medical complications, permanent disability or even death. Patients in ICUs are particularly vulnerable to infections because the body's ability to fight infection is reduced by illness or trauma. Surgical tubes and lines (for example to help with feeding or breathing) may enable bacteria to enter the body. Chlorhexidine is a low-cost product which is used as an antiseptic and disinfectant in hospitals.

What are the main results of the review?

In December 2018 we searched for studies looking at the use of chlorhexidine for bathing critically ill patients. We found eight studies dating from 2005 to 2018, involving a total of 24,472 people across more than 20 ICUs. Seven studies included people who were adults, and one study included only children. All studies included both males and females. All studies compared bathing with chlorhexidine versus bathing with soap and water or non-antimicrobial washcloths. Four studies received funding from independent funders (government organisations, or from hospital or university departments) or reported no external funding, and four studies received funding from companies that manufactured chlorhexidine products.

The evidence from all eight studies combined is not sufficient to allow us to be certain whether patients bathed in chlorhexidine are less likely to catch an infection during their stay in the ICU. We are also uncertain whether patients bathed in chlorhexidine are less likely to die, because the certainty of the evidence from the six studies that reported on this is very low. We did not pool the evidence from the six studies that reported how long patients had stayed in the ICU, because the results differed widely. We are also uncertain whether patients bathed in chlorhexidine are likely to be in the ICU for less time, because the certainty of the evidence is very low. Reports from five studies provided different evidence about whether chlorhexidine led to more or less skin reactions; we are uncertain whether patients bathed in chlorhexidine are likely to have more or less skin reactions, because the certainty of the evidence is very low.

Quality of evidence

Most studies did not use methods to conceal the type of bathing solution that staff were using, which increases the risk that staff may have treated patients differently depending on whether patients were in the chlorhexidine study group or the soap-and-water study group. Participants in some studies may have already caught an infection before the start of the study and we were concerned that this might have affected our results. We also noticed wide differences in some results, and some outcomes had few reported events. These were reasons to judge the quality of the evidence to be very low.

How up to date is this review?

We searched for studies that had been published up to December 2018.

SUMMARY OF FINDINGS

Summary of findings for the main comparison. Bathing of the critically ill with chlorhexidine versus bathing with soap and water or non-antimicrobial washcloths for the prevention of hospital acquired infections

Bathing of the critically ill with chlorhexidine versus bathing with soap and water or non-antimicrobial washcloths for the prevention of hospital acquired infections

Population: people who are critically ill

Settings: ICUs in France, Italy, Thailand, and USA; studies included single-centre or multicentre settings

Intervention: bathing with a solution of chlorhexidine versus bathing with a solution of soap and water or non-antimicrobial washcloths

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	Number of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Assumed risk with soap and water bathing	Assumed risk with chlorhexidine bathing				
Hospital-acquired infections Data collected during ICU stay	Study population		Rate difference 1.70 (0.12 to 3.29)	21,924 (8 studies)	⊕⊕⊕⊕ Very low^a	We are uncertain whether using chlorhexidine for bathing of critically-ill people reduced the rate of hospital-acquired infection. We used data from cluster-randomised cross-over studies in which appropriate adjustments were made for study design. We calculated rate difference using generic inverse variance in order to account for studies that reported data as number of events or rates.
	9.5 infections per 1000 patient days	7.8 infections per 1000 patient days (6.2 to 9.4)				
Mortality Data collected (where reported) in hospital, in the ICU, and at 48 hours	Study population		OR 0.87 (0.76 to 0.99)	15,798 (6 studies)	⊕⊕⊕⊕ Very low^b	We are uncertain whether using chlorhexidine for bathing of critically-ill people reduced mortality. We used standard errors imputed using an estimated design effect for 2 cluster-randomised cross-over studies. We calculated OR using generic inverse variance.
	9.7 deaths per 100 patients	8.5 deaths per 100 patients (7.6 to 9.6)				
Length of stay in the ICU	Study population		Not estimable	18,570 (6 studies)	⊕⊕⊕⊕ Very low^c	We are uncertain whether using chlorhexidine for bathing of critically-ill people reduced length of stay in the ICU. We did not conduct meta-analysis because data were skewed. We noted no evidence of any difference in effect in each study.
	7 days (median)	Not estimable				
Adverse effects: skin reactions. Re-	Of participants bathed with chlorhexidine, 1 study reported 5 mild skin reaction, 1 study reported 1 mild skin reaction, 1 study report-		Not estimable	6365 (5 studies)	⊕⊕⊕⊕ Very low^d	We are uncertain whether using chlorhexidine for bathing of critically-ill people reduced adverse events.

We did not combine data due to insufficient information from study authors or incomparable data. Two additional studies reported skin reactions but believed that these were not attributable to the bathing solution.

ported as attributable to chlorhexidine or soap and water.

Data collected during ICU stay

ed 12 skin reactions, and 1 study reported 21 skin reactions. Comparative data for the control was not clearly reported in 2 studies and 1 study reported 23 skin reactions, respectively. In 1 multi-armed study, 6 participants in 2 chlorhexidine groups and 6 participants in 2 control groups had skin reactions

*The basis for the **assumed risk** is the median control group risk across studies. The **corresponding risk** (and its 95% CI) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: confidence interval; OR: odds ratio; ICU: intensive care unit

GRADE Working Group grades of evidence

High certainty: further research is very unlikely to change our confidence in the estimate of effect.

Moderate certainty: further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.

Low certainty: further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.

Very low certainty: we are very uncertain about the estimate.

^aDowngraded by three levels: one level for study limitations (high risk of performance bias in most studies, high risks of other bias in individual studies); one level for inconsistency (sensitivity analysis showed that one small study had a large influence on result, and use of an alternative design effect in one cluster-randomised cross-over study changed the effect); one level for indirectness (participants in some studies may have had infections before randomisation)

^bDowngraded by three levels: one level for study limitations (high risk of performance bias in most studies, high risks of other bias in individual studies); one level for inconsistency (sensitivity analysis showed that use of an alternative design effect in two cluster-randomised cross-over studies changed the effect); one level for indirectness (participants in some studies may have had infections before randomisation)

^cDowngraded by three levels: one level for study limitations (high risk of performance bias in most studies, high risks of other bias in individual studies); one level for imprecision (visual inspection of data showed skewed data); one level for indirectness (participants in some studies may have had infections before randomisation)

^dDowngraded by three levels: one level for study limitations (high risk of performance bias in most studies, high risks of other bias in individual studies); one level for imprecision (events are very few); one level for indirectness (participants in some studies may have had infections before randomisation)

BACKGROUND

Description of the condition

Hospital-acquired infection is one of the most frequent types of adverse event to affect patient care, and can lead not only to discomfort and increased length of stay in hospital, but also to permanent disability and even death. The prevalence of such infections varies internationally, and there are limited data from low-income countries, where the rates are greater than in high-income countries. Examples of prevalence include up to 6% of patients in the UK (Health Protection Agency 2016), and 4% of patients in the USA (Magill 2014), whilst reports of prevalence in settings with limited resources vary, for example 5.4% in Mongolia (Ider 2010), 14.5% in Tunisia (Mahjoub 2015) and 19.1% in Albania (Faria 2007).

Whilst all people staying in hospital are susceptible to infections, prevalence in the intensive care unit (ICU) is particularly high. A one-day prospective, multi-centre, international study reported 51% of adult patients were classified as infected, and the rate of infection increased to more than 70% for people whose ICU stay was seven days or longer (Vincent 2009). Patients in ICUs are critically ill; they have suppressed immunity as a result of trauma, injury or blood loss (or a combination of these), which increases their susceptibility to infection (Volk 2002). In addition, people who are mechanically-ventilated in the ICU are at risk due to tracheostomy, reintubation and the use of multiple central venous catheters (Ibrahim 2001), where lines and tubes may act as a vector for the transmission of bacteria and lead to ventilator-associated pneumonia (VAP).

Common pathogens in hospital-acquired infection include *Staphylococcus aureus*, *Clostridium difficile* and *Enterococci*; and the overuse of broad spectrum antibiotics has promoted bacteria which are drug-resistant and difficult to treat (Bereket 2012). Methicillin-resistant *S. aureus* (MRSA) causes a range of infections including abscesses, surgical site infections, gastroenteritis, pneumonia, urinary tract infections and endocarditis. It is transmitted by direct contact with an infected person or their environment (or both), and colonises the skin or nostrils. Similarly transmitted, vancomycin-resistant *Enterococci* (VRE) leads to urinary tract infections, skin/wound infections, and intra-abdominal infections. *C. difficile* causes diarrhoea following administration of antibiotics, and is transmitted through the faecal-oral route by an infected person or environment (Kelly 2012).

In 2009, Vincent and colleagues reported the most common sites of infection in the ICU as the respiratory tract, abdominal, bloodstream and renal/urinary tract, with respiratory tract infections representing 63.5% of these (Vincent 2009). Healthcare packages and guidelines are now being established to reduce hospital-acquired infections and subsequent morbidity and mortality rates, for example, a 'central line bundle' of care is being used to try to reduce central line-associated bloodstream infections (CLABSI), which includes interventions such as education programmes for personnel, hand hygiene and daily review of the need for catheters (Sacks 2014).

Description of the intervention

Chlorhexidine is a biocide on the World Health Organization's List of Essential Medicines (WHO 2017). It has a broad spectrum of action, destabilising the cell walls of gram positive and gram

negative bacteria and fungi (Puig Silla 2008; WHO 2011). It can kill most bacteria within 30 seconds of contact (Genuit 2001). In binding to proteins in human tissue, such as skin and mucous membranes, chlorhexidine can also have a slow-release action, with prolonged activity up to 48 hours after the initial application (Hibbard 2005), and this residual antibacterial activity suggests that organisms that come into contact after chlorhexidine use may not be able to grow (Wade 1991). Chlorhexidine is known to be effective against organisms present in hospital-acquired infections including *S. aureus* and *Enterococcus* (McDonnell 1999).

Chlorhexidine is widely used as a disinfectant and antiseptic in applications such as oral hygiene mouthwashes, hand disinfectants, wound cleansers and preoperative skin preparation (McDonnell 1999). Concentrations range from 0.004% to 4%, in alcohol or aqueous pharmaceutical solution, and it is available in these different dose forms as gels, lotions, solutions, and liquids, and in pads, dressings and sponges.

Chlorhexidine is a low-cost product. Cochrane systematic reviews have demonstrated that it is effective in particular situations, for example in the reduction of neonatal mortality when used for skin and umbilical cord care in the community setting (Sinha 2015), and in the reduction in rates of ventilator-acquired pneumonia when used in dental hygiene care of people in the ICU (Shi 2013).

How the intervention might work

People in ICUs are subject to increased invasive monitoring by healthcare personnel. They may be mechanically-ventilated, have central venous catheters, arterial lines, intravenous catheters, urinary catheters and/or chest tubes, as well as having wounds (both surgical and trauma). All these factors increase the risk of transmission of infection in people who also have reduced immunity (Inweregbu 2005).

Using an antibacterial solution that disinfects the whole skin area during bathing of part or all of the body, may quickly begin to kill existing bacteria. However, chlorhexidine may also form a 'protective coating' to further reduce the risk of hospital-acquired infections, such as VAP, CLABSI, catheter-related blood stream infections (CRBSI) and catheter-associated urinary tract infection (CAUTI) in this high-risk population.

Although chlorhexidine is known to be a low-risk skin irritant, the risk of irritation, such as contact dermatitis, may differ between chlorhexidine products with differing concentrations (Calogiuri 2013; McDonnell 1999).

Why it is important to do this review

Hospital-acquired infections are estimated to lead to 37,000 deaths in Europe, with additional financial burdens (for example through prolonged hospital stay) of EUR 7 billion a year, and up to 99,000 annual deaths in the USA and costs of USD 6.5 billion (WHO 2011).

Morbidity and mortality related to such infections is preventable. People in the ICU are inevitably at high risk, and establishing strategies to reduce rates of infection (in this case, establishing the effectiveness of bathing with a suitable solution) would be beneficial to healthcare systems worldwide, improving outcomes for people who stay in hospital and reducing the length of hospital and ICU stay.

As yet, there are no reports of chlorhexidine-resistant bacteria. However, chlorhexidine is a widely used product and there are reports of reduced susceptibility of MRSA to chlorhexidine (Horner 2012).

It is important to assess the potential benefits and harms of chlorhexidine for bathing people who stay in the ICU.

OBJECTIVES

To assess the effects of chlorhexidine bathing on the number of hospital-acquired infections in people who are critically ill.

METHODS

Criteria for considering studies for this review

Types of studies

We included randomised controlled trials (RCTs). We included both parallel and cross-over designs, as well as cluster and non-cluster designs. We only included cross-over designs if data was available for the participants or clusters randomised to the initial treatment group.

Types of participants

We included adult and child participants with any condition that required admission to the intensive care unit (ICU). We included admission to high-dependency or critical care units or other hospital wards specifically designed to cater for people who are critically ill. We did not include studies of neonates.

We had intended to exclude studies in which participants were diagnosed with a hospital-acquired infection prior to randomisation, but we found that this was not clearly reported in studies. We therefore noted how this was reported in each included study and considered it during the 'Risk of bias' assessment. See [Differences between protocol and review](#).

Types of interventions

We included studies that compared bathing with a solution of chlorhexidine by any means (e.g. impregnated washcloths or chlorhexidine gel) to bathing using an alternative solution (e.g. soap and water) or no bathing. We defined bathing as the washing of all body areas either at the bedside (e.g. wipe with an impregnated cloth) or in a bath or shower; we excluded studies in which only one body area was washed with a solution of chlorhexidine. We included studies of bathing interventions at different frequencies, for example daily washing or weekly washing.

Types of outcome measures

Our primary interest was whether bathing with chlorhexidine reduced the risk of any hospital-acquired infection and we therefore recorded the number of participants who acquired an infection since the introduction of the intervention. We included data that were collected from appropriate clinical evaluation of symptoms, or physical signs of infection, or laboratory test results. We collected mortality data from any cause. We collected data for the length of stay in the ICU as number of days. We recorded the number of participants who had any reaction that may be attributable to the intervention or comparison (to include known adverse effects such as skin irritation, rash, contact dermatitis,

redness, blistering, swelling of face, hands or feet, or difficulty breathing).

Primary outcomes

1. Hospital-acquired infections, including bloodstream infections; central-line associated bloodstream infections; ventilator-associated pneumonia; catheter-associated urinary tract infections; multidrug-resistant organisms (MDROs), e.g. Methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant *Enterococci* (VRE).

Secondary outcomes

1. Mortality.
2. Length of stay in the ICU.
3. Adverse effects, including skin irritation, or responses such as swelling of face, hands or feet, or breathing difficulties (as defined by the study authors).

Search methods for identification of studies

Electronic searches

We searched the following electronic databases to identify reports of RCTs:

- The Cochrane Wounds Specialised Register (searched 10 December 2018);
- The Cochrane Central Register of Controlled Trials (CENTRAL; 2018, Issue 11) in the Cochrane Library (searched 10 December 2018);
- Ovid MEDLINE (1946 to 10 December 2018);
- Ovid Embase (1974 to 10 December 2018);
- EBSCO CINAHL Plus (Cumulative Index to Nursing and Allied Health Literature; 1937 to 10 December 2018).

The search strategies for the Cochrane Wounds Specialised Register, CENTRAL, Ovid MEDLINE, Ovid Embase and EBSCO CINAHL Plus can be found in [Appendix 1](#). We combined the Ovid MEDLINE search with the Cochrane Highly Sensitive Search Strategy for identifying randomised trials in MEDLINE: sensitivity- and precision-maximising version (2008 revision) (Lefebvre 2011). We combined the Embase search with the Ovid Embase filter terms developed by the UK Cochrane Centre (Lefebvre 2011). We combined the CINAHL Plus searches with the trial filters developed by the Scottish Intercollegiate Guidelines Network (SIGN 2018). There were no restrictions with respect to language, date of publication or study setting.

We also searched the following clinical trial registries:

- ClinicalTrials.gov (www.ClinicalTrials.gov/) (searched 10 December 2018)([Appendix 2](#));
- World Health Organization (WHO) International Clinical Trials Registry Platform (www.who.int/ictrp/search/en/) (searched 10 December 2018)([Appendix 3](#)).

Searching other resources

We carried out backward citation searching of key reviews identified from the searches. We carried out forward citation searching of included studies.

We also carried out grey literature searching through 'Opengrey' (www.opengrey.eu/).

Data collection and analysis

We carried out data collection and analysis according to the methods stated in the published protocol (Lewis 2016), which were based on the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins 2011).

Two review authors (Sharon Lewis (SL) and Oliver Schofield-Robinson (OSR)) independently carried out all initial data collection and analysis, before comparing results and reaching consensus. A third author was available to resolve conflicts if required. An additional author (Sarah Rhodes (SR)) was introduced after data extraction to help incorporate cluster-randomised cross-over trials into the analysis.

Selection of studies

We used reference management software to collate the results of the searches and to remove duplicates (Endnote 2011).

Two review authors (SL and OSR) used Covidence 2017 software to screen the results of the search from the titles and abstracts and identify any potentially relevant studies from this information alone. Two review authors (SL and OSR) sourced the full texts of all those potentially relevant studies and considered whether they met the inclusion criteria (see [Criteria for considering studies for this review](#)). We planned to include abstracts at this stage if they contained sufficient information and relevant results that included denominator figures for each intervention/comparison group.

We recorded the number of papers retrieved at each stage and reported this using a PRISMA flow chart (Liberati 2009). We collected brief details of closely related but excluded papers.

Data extraction and management

Two review authors (SL and OSR) used Covidence 2017 to extract data from individual studies. We extracted the following information.

1. Methods: type of study design; setting; dates of study; funding sources.
2. Participants: number of participants randomised to each group; baseline characteristics (including Acute Physiology and Chronic Health Evaluation II (APACHE II) scores).
3. Interventions: details of intervention and comparison, including concentration of chlorhexidine.
4. Outcomes: review outcomes measured and reported by study authors.
5. Outcome data: results of outcome measures.

We considered the applicability of information from individual studies and generalisability of the data to our intended study population (i.e. the potential for indirectness in our review).

There were multiple publications of some studies. In this case, we created a composite data set from all the eligible publications.

Assessment of risk of bias in included studies

We assessed study quality, study limitations and the extent of potential bias using the Cochrane 'Risk of bias' tool (Higgins 2017). See [Appendix 4](#). We considered the following domains.

1. Sequence generation (selection bias);
2. Allocation concealment (selection bias);
3. Blinding of participants, personnel and outcome assessors (performance and detection bias);
4. Incomplete outcome data (attrition bias);
5. Selective outcome reporting (reporting bias);
6. Other potential risks of bias: use of concomitant methods to reduce infection.

We anticipated that there would be a risk of performance bias in the methodology of the studies included in this review, and we noted any methods used by study authors to minimise this risk. We expected that robust study methodology would include blinding of outcome assessors as some outcomes could be measured at a later stage and by personnel not involved with the bathing routine. We anticipated that different hospital units were likely to follow different practices for infection prevention and control in addition to bathing, e.g. use of antiseptic or antibiotic-coated catheters. We collected available data of any additional infection prevention strategies and noted whether these were likely to be equivalent between groups.

For cluster-randomised cross-over study designs, we referred to particular guidance on assessing risk of bias in cluster-randomised studies and in cross-over studies (Higgins 2011) (see [Appendix 5](#)). In particular, we assessed: recruitment bias; loss of clusters; baseline imbalances between clusters; and whether analysis was appropriate for the cluster design.

For each domain, we judged whether study authors had made sufficient attempts to reduce bias. We made our judgements using one of three measures (low risk, high risk, unclear). We recorded this in 'Risk of bias' tables and present a summary 'Risk of bias' figures.

Measures of treatment effect

We recorded the number of hospital-acquired infections as rate differences; this was a change from the original protocol (see [Differences between protocol and review](#)). Mortality was recorded as dichotomous data in order to calculate odds ratios (OR), and we reported the number of adverse events as dichotomous data.

We recorded length of stay as continuous data.

Unit of analysis issues

We identified one study which had a 2x2 factorial design (Camus 2005). Only one arm included chlorhexidine and we selected this intervention arm (which also included mupirocin as a treatment agent) and compared it to the group with no active agent. We did not include any multi-armed studies comparing more than one type of chlorhexidine bathing.

In this review, we encountered studies that were randomised by cluster and also included a cross-over design. For studies that used analysis methods to take account of both the clustering effect and the cross-over design, we extracted appropriately adjusted

standard errors (SEs) for meta-analysis using the generic inverse-variance method. For studies in which appropriate adjusted SEs were not reported, we applied appropriate adjustment using an estimate of the design effect for each study (Higgins 2011).

The standard formulae to calculate the design effect of cluster-randomised studies only takes into account the effect of clustering (which we would expect to increase the SE), but not the effect of the cross-over design (which we would expect to reduce the SE) (Higgins 2011). We aimed to estimate the square root of the design effect for cluster-randomised cross-over studies as (unadjusted SE)/(adjusted SE) when we could obtain crude SEs and SEs that adjusted for clustering and cross-over design. This estimation method assumes that the design effect is consistent across each outcome in the same study; therefore, when this estimation method was used, we interpreted the results with caution.

Please see [Differences between protocol and review](#) for details of changes to this section.

Dealing with missing data

We contacted study authors to clarify missing data. We used available reported data if necessary, rather than imputing values.

Assessment of heterogeneity

We assessed whether there was evidence of inconsistency within our results through consideration of heterogeneity. We assessed clinical heterogeneity by comparing similarities between the participants, interventions and outcomes in the included studies. We assessed statistical heterogeneity by calculation of the Chi^2 (with an associated P value) or I^2 measure (with an associated percentage). We used the following values as a guide to interpretation: I^2 at 0% to 40% is not considered important, 30% to 60% suggests moderate heterogeneity, 50% to 90% suggests substantial heterogeneity, and 75% to 100% represents considerable heterogeneity (Higgins 2011). When assessing heterogeneity, we also considered the point estimates and the overlap of confidence intervals (CIs). If the CIs overlapped then we considered the results to be more consistent. However, it is possible for combined studies to show a large consistent effect but with significant heterogeneity. We therefore interpreted heterogeneity with caution (Guyatt 2011b).

Assessment of reporting biases

We attempted to source published protocols for each of our included studies using clinical trial registers. We compared published protocols with published study results, to assess the risk of selective reporting bias.

We did not have sufficient studies, i.e. more than 10 (Sterne 2017), to generate a funnel plot to assess the risk of publication bias in the review. An asymmetric funnel plot may indicate the publication of only positive results (Egger 1997).

Data synthesis

We completed meta-analysis for outcomes where comparable effect measures were available from more than one study, and where measures of heterogeneity indicated that pooling of results was appropriate.

For hospital-acquired infections, we analysed rate differences by entering the rate difference and the associated SE into the generic inverse variance function in [Review Manager 2014](#). This method accounted for the inclusion of cluster-randomised cross-over studies. For mortality, we used generic inverse variance to calculate the log OR, which also accounted for the inclusion of cluster-randomised cross-over studies. We used a random-effects model in all analyses to account for the anticipated differences in illness severity or participant conditions. For length of stay in the ICU we planned to use mean difference, and for adverse events we planned to use the OR. See [Differences between protocol and review](#).

We calculated CIs at 95% and used a P value of 0.05 or less to judge whether a result was statistically significant.

We considered whether there was imprecision in the results of analyses by assessing the CI around an effects measure; a wide CI would suggest a higher level of imprecision in the results. A small number of studies may also reduce the precision (Guyatt 2011a).

Subgroup analysis and investigation of heterogeneity

We did not identify sufficient studies to explore differences between them using subgroup analysis. If there had been more than 10 studies (Deeks 2017), we would have conducted subgroup analyses for the following:

1. illness severity (e.g. based on APACHE II scores);
2. age of participants (e.g. infants, adults, older adults);
3. invasive device use (e.g. intravascular devices, mechanical ventilation, feeding lines).

Sensitivity analysis

We explored the potential effects of decisions made as part of the review process as follows:

1. we excluded all studies that we judged to be at high or unclear risk of selection bias;
2. we excluded studies in which participant outcome data were missing, for which we used available reported data;
3. we conducted meta-analysis using the alternate meta-analytic effects model (fixed-effect versus random-effects).

We compared effect estimates from the analysis of our primary outcome with effect estimates calculated during the above sensitivity analyses. We reported differences that altered our interpretation of the effect.

In addition to planned sensitivity analyses, we considered the effect of including cluster-randomised cross-over study designs in the review. We imputed more conservative SEs using standard adjustment for clustering, but ignored the effect of the cross-over design. See [Differences between protocol and review](#).

'Summary of findings' tables

The GRADE approach incorporates assessment of indirectness, study limitations, inconsistency, publication bias and imprecision (GRADE 2013). We used the assessments made during our analysis to inform the GRADE process (see [Data extraction and management](#), [Assessment of risk of bias in included studies](#), [Assessment of heterogeneity](#), [Assessment of reporting biases](#)

and [Data synthesis](#), respectively). This approach gives an overall measure of how confident we can be that our estimate of effect is correct ([Guyatt 2008](#)).

We used the principles of the GRADE system to give an overall assessment of the evidence relating to each of the following outcomes:

1. hospital-acquired infections;
2. mortality;
3. length of stay;
4. adverse event: skin irritation.

Two authors (SL and OSR) independently used the GRADEpro Guideline Development Tool software to create a 'Summary of findings' table ([GRADEpro 2015](#)). We assessed the evidence for limitations, inconsistency, indirectness, publication bias and imprecision using the following ratings of certainty: high; moderate; low and very low. We reached consensus and resolved disagreements through informal discussion, with a third review author available if further consultation had been required.

RESULTS

Description of studies

Results of the search

We screened 532 titles and abstracts from database searches, and sourced the full text of 56 potentially eligible studies. Of these, we identified 12 records of 8 studies that were eligible for inclusion in our review. There were multiple publications of some studies and we combined these into eight unique studies.

We identified ten reviews from the database searches ([Afonso 2013](#); [Afonso 2016](#); [Chen 2015](#); [Choi 2015](#); [Derde 2012](#); [Frost 2016](#); [Huang 2016](#); [Kim 2016](#); [O'Horo 2012](#); [Shah 2016](#)). We carried out backward citation searching on these and did not identify any additional studies for inclusion. We carried out forward citation tracking on our eight included studies using Google Scholar and Web of Science, and identified no additional studies eligible for inclusion.

We also carried out searches of clinical trial registers and identified clinical trial reports for seven of our included studies. From this search, we found one completed study without published results, and two ongoing studies. We carried out a grey literature search and found no studies that matched our criteria. See [Figure 1](#).

Figure 1. Flow diagram

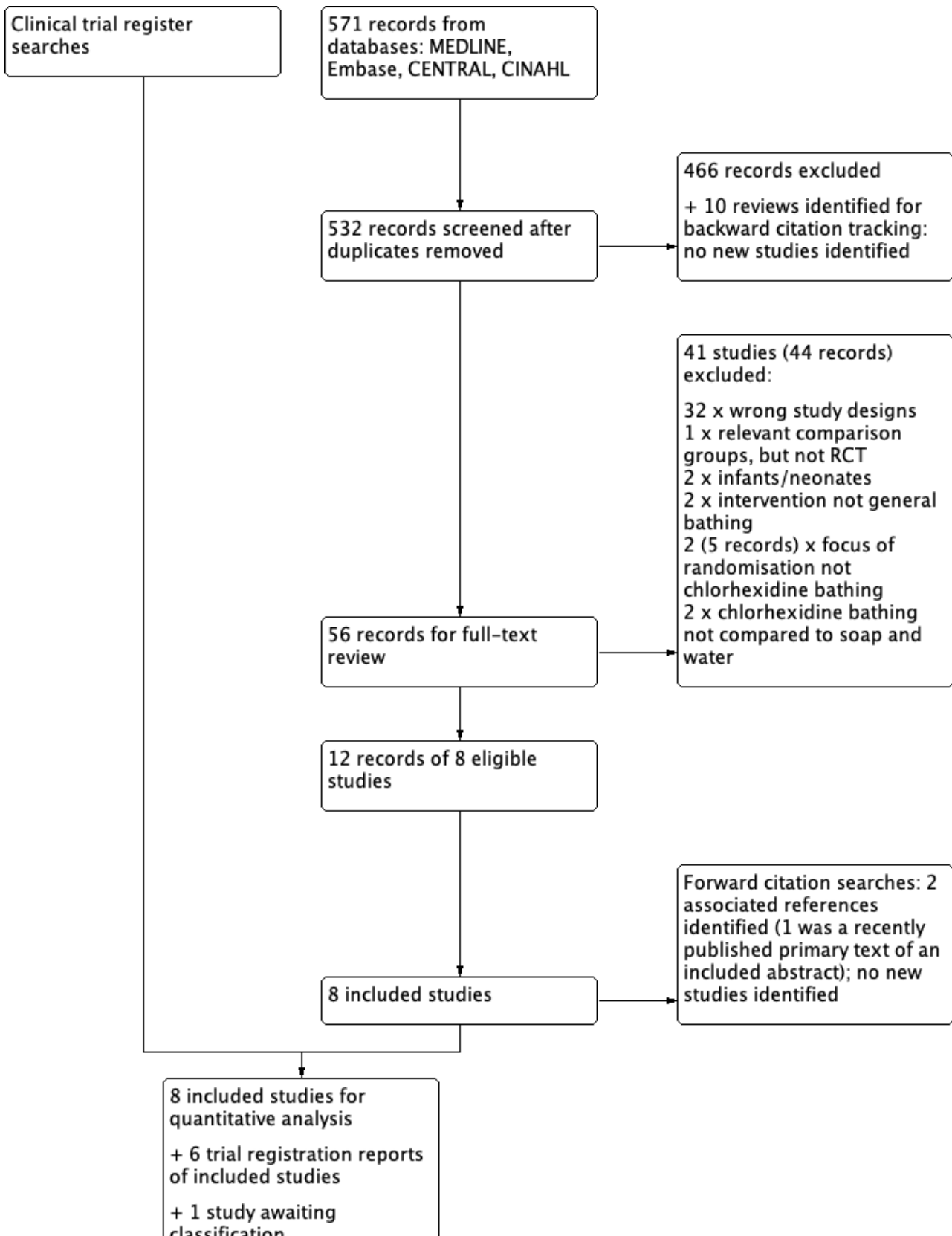


Figure 1. (Continued)

+ 1 study awaiting classification

+ 2 ongoing studies

Included studies

See [Characteristics of included studies](#).

Types of studies

We included eight studies ([Bleasdale 2007](#); [Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)). Four studies were randomised controlled trials (RCTs) and included 1537 individually randomised participants ([Boonyasiri 2016](#); [Camus 2005](#); [Pallotto 2018](#); [Swan 2016](#)); four studies were cluster-randomised cross-over studies with the ICU as the unit of randomisation, and they included 23 randomised ICUs with 22,935 participants ([Bleasdale 2007](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)).

Types of participants and setting

Four studies were conducted within a single centre ([Bleasdale 2007](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)), and four were conducted in multiple centres ([Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#)). The ICUs in which the studies were conducted were general, medical, surgical, trauma, neurological, cardiac care, and respiratory care. All studies included adult participants except [Milstone 2013](#), which included only paediatric participants.

Five studies did not report whether any participants had a hospital-acquired infection at enrolment ([Bleasdale 2007](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#); [Pallotto 2018](#)). Two studies reported that some participants had infections prior to randomisation ([Camus 2005](#); [Swan 2016](#)); we have reported the number of infections with the respective study baseline characteristics ([Characteristics of included studies](#)) One study did not report hospital-acquired infections at baseline but reported multi-drug-resistant bacteria colonisation, which was balanced between groups ([Boonyasiri 2016](#)).

Types of interventions and comparisons

Five studies compared daily bathing using 2% chlorhexidine-impregnated washcloths, with daily or twice daily soap-and-water bathing or bathing with non-antimicrobial washcloths ([Bleasdale 2007](#); [Boonyasiri 2016](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)). One study compared alternate-day bathing using washcloths submerged in a solution of 4% chlorhexidine, diluted with warm water to 2%, with soap-and-water bathing or bathing with washcloths ([Swan 2016](#)). One study compared once-daily bathing with 4% chlorhexidine using washcloths followed by water rinsing ([Pallotto 2018](#)). Another study used 4% chlorhexidine at a 12-hourly rate, compared with liquid soap; there were no further details of dilution or bathing methods in this study ([Camus 2005](#)). Camus and colleagues employed a 2 x 2 factorial design in which chlorhexidine was combined with mupirocin to form one intervention, which was compared with another intervention group (polymyxin and tobramycin) and two control groups ([Camus 2005](#)); we included data for the chlorhexidine and mupirocin group, compared to a control group that did not have any active

intervention. Impregnated washcloths were pre-manufactured by pharmaceutical companies in four studies ([Bleasdale 2007](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)) and prepared by the hospital pharmacy in one study ([Boonyasiri 2016](#)); this information was not reported in one study ([Pallotto 2018](#)).

Outcomes

We collected data for hospital-acquired infections from eight studies ([Bleasdale 2007](#); [Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)). Six studies reported mortality data ([Boonyasiri 2016](#); [Camus 2005](#); [Milstone 2013](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)), and six studies reported the length of stay in ICU ([Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)). Adverse effects of skin irritation were reported in seven studies ([Bleasdale 2007](#); [Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Pallotto 2018](#); [Swan 2016](#)). Other adverse effects were not reported.

Funding sources

Three studies received institutional funding ([Boonyasiri 2016](#); [Noto 2015](#); [Swan 2016](#)) and one study reported that no external funding was received ([Pallotto 2018](#)); and four studies reported full or partial funding from companies which manufacture chlorhexidine products ([Bleasdale 2007](#); [Camus 2005](#); [Milstone 2013](#); [Noto 2015](#)).

Excluded studies

We excluded 41 (44 reports) studies at the stage of full-text review (see [Figure 1](#)). We excluded 32 studies (with 32 reports) that were the wrong study design (i.e. editorials, letters/comments, reviews, and study designs that were not RCTs. See [Appendix 6](#)). We did not report details of these 32 studies in the review. In addition, we excluded nine RCTs (with 12 reports) and we report details of these key studies in [Characteristics of excluded studies](#). Two studies had used chlorhexidine bathing with newborn infants and we believed that these were not comparable with studies of a general ICU population ([Cunha 2008](#); [Sankar 2009](#)). One study randomised participants specifically for bathing of the perineal area to prevent catheter-associated urinary tract infections (CAUTIs), and was not comparable with studies of general bathing ([Choi 2012](#)). One study compared solutions used to cleanse the periurethral area prior to urinary catheter placement ([Duzkaya 2017](#)). One study randomised participants to receive screening for Methicillin-resistant *Staphylococcus aureus* (MRSA) and only administered chlorhexidine bathing to those within the intervention group who were MRSA-positive ([Camus 2011](#)). We excluded one study which was both an interrupted-time series and an RCT, however the focus of randomisation was on screening, rather than chlorhexidine use ([Derde 2014](#)); we identified two associated conference abstract references for this study. One RCT included a relevant intervention group of chlorhexidine bathing, however the comparison was screening, isolation and decolonization strategies, not soap-and-water bathing or no bathing ([Huang 2013](#); we identified one associated reference to this study. Another study compared two

different methods of chlorhexidine bathing, and did not employ comparison groups of soap-and-water bathing or no bathing (Dean 2011). One study compared chlorhexidine bathing with soap and water, in a prospective cross-over study, but it was not randomised (Lowe 2017). See [Characteristics of excluded studies](#).

Studies awaiting classification

We identified one study that was registered with a clinical trial register and described as having completed participant recruitment (ChiCTR-TRC-13004164). We have been unable to source a report of this study and have contacted the authors to request information. We are awaiting any relevant information. See [Characteristics of studies awaiting classification](#).

Ongoing studies

We identified two ongoing studies (IRCT2017030932293N1; NCT02870062). Both RCTs include use of daily chlorhexidine bathing with adults in the ICU. The anticipated recruitment is 80 participants (IRCT2017030932293N1), and 40 participants (NCT02870062).

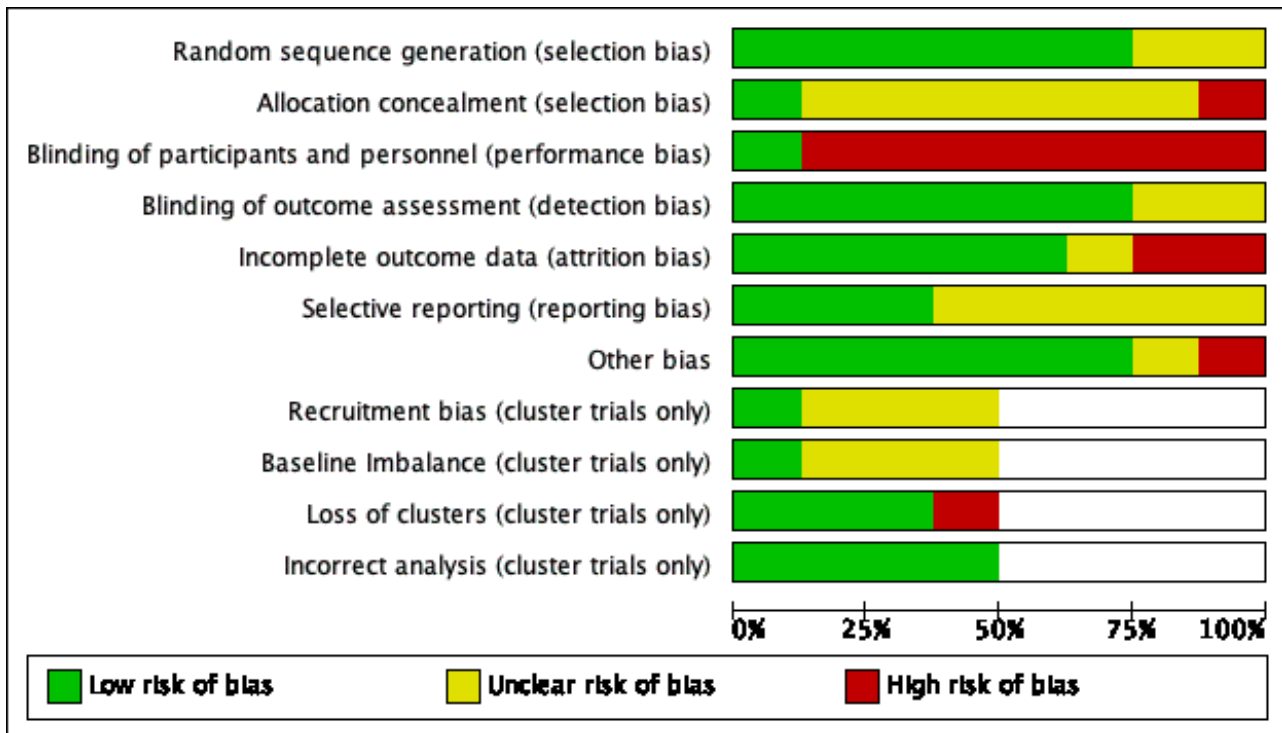
Risk of bias in included studies

For a summary of the 'Risk of bias' assessments, see [Figure 2](#) and [Figure 3](#).

Figure 2. Risk of bias summary: review authors' judgements about each risk of bias item for each included study.

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Blinding of participants and personnel (performance bias)	Blinding of outcome assessment (detection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Other bias	Recruitment bias (cluster trials only)	Baseline imbalance (cluster trials only)	Loss of clusters (cluster trials only)	Incorrect analysis (cluster trials only)
Bleasdale 2007	?	-	-	+	+	+	+	+	+	+	+
Boonyasiri 2016	+	+	-	+	-	+	+				
Camus 2005	+	?	+	?	+	?	-				
Climo 2013	?	?	-	?	-	?	?	?	?	-	+
Milstone 2013	+	?	-	+	?	?	+	?	?	+	+
Noto 2015	+	?	-	+	+	?	+	?	?	+	+
Pallotto 2018	+	?	-	+	+	?	+				
Swan 2016	+	?	-	+	+	+	+				

Figure 3. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies.



Allocation

All eight included studies were described as randomised trials. We judged four RCTs to be at low risk of selection bias (Boonyasiri 2016; Camus 2005; Pallotto 2018; Swan 2016): all had reported adequate methods of randomisation. For the cluster-randomised cross-over studies, each study used a separate ICU for each cluster, and randomisation was completed at cluster level. Two studies reported adequate methods of randomisation; we judged these studies to have low risk of bias (Milstone 2013; Noto 2015). We could not be certain of the risk of bias in two studies because methods were not described (Bleasdale 2007; Climo 2013).

One study reported adequate methods to conceal allocation (Boonyasiri 2016). Four studies reported no methods to conceal allocation and we assessed the risk of bias as unclear (Camus 2005; Climo 2013; Noto 2015; Pallotto 2018). Two studies did not provide adequate information for allocation concealment and we judged these to also have an unclear risk of bias (Milstone 2013; Swan 2016). One study had only two clusters; we believed that allocation concealment was not feasible and the risk of bias was high (Bleasdale 2007).

Blinding

Blinding of participants and hospital personnel was not undertaken in seven studies and we judged these to have a high risk of performance bias (Bleasdale 2007; Boonyasiri 2016; Climo 2013; Milstone 2013; Noto 2015; Pallotto 2018; Swan 2016). Only one study described adequate methods to blind both the participants and the personnel to the intervention (Camus 2005), and we judged this to have a low risk of performance bias. Six studies reported that outcome assessors were blinded to group allocation (Bleasdale 2007; Boonyasiri 2016; Milstone 2013; Noto 2015; Pallotto 2018;

Swan 2016). Two studies did not provide sufficient detail of whether outcome assessors were blinded and we judged the detection bias as representing an unclear risk (Camus 2005; Climo 2013).

Incomplete outcome data

We assessed studies that reported no losses or few losses as having low risk of attrition bias (Bleasdale 2007; Camus 2005; Noto 2015; Pallotto 2018; Swan 2016). One study had a large number of losses in one group because of lack of consent; study authors used an intention-to-treat (ITT) analysis and reported that this was comparable to a per-protocol analysis (Milstone 2013); we assessed this study as having unclear risk of attrition bias. We noted a large number of losses in Boonyasiri 2016, which may have influenced the results of this study. In Climo 2013, study authors reported no losses and had used an ITT analysis but we noted discrepancies in the reported number of randomised participants. We judged two studies to have a high risk of attrition bias (Boonyasiri 2016; Climo 2013).

Selective reporting

Five studies were prospectively registered with clinical trial registers (Bleasdale 2007; Boonyasiri 2016; Climo 2013; Milstone 2013; Swan 2016); three of these had reported outcomes in the final report which matched those in the clinical trial register documents and we judged these studies to have low risk of reporting bias (Bleasdale 2007; Boonyasiri 2016; Swan 2016). Two had inconsistencies between outcomes listed in the clinical trial register documents and the final report, and we were unclear if this introduced bias (Climo 2013; Milstone 2013). Two studies were retrospectively registered with a clinical trial register (Noto 2015; Pallotto 2018), and we were unable to identify clinical trial

registration for the remaining study (Camus 2005). It was therefore not feasible to judge any risks of reporting bias for these studies.

Other potential sources of bias

We identified no additional sources of bias in six studies (Bleasdale 2007; Boonyasiri 2016; Milstone 2013; Noto 2015; Pallotto 2018; Swan 2016). We noted a lack of wash-out period in Climo 2013, but we judged that the study investigators had addressed this risk effectively. We judged the study by Camus and colleagues to have a high risk of bias because the chlorhexidine group also included treatment with mupirocin, which was not given to the participants in the control group (Camus 2005). We also noted that more participants in the control group in this study had a hospital-acquired infection prior to randomisation, which introduced a high risk of bias.

Recruitment bias (cluster trials only)

We judged the risk of recruitment bias to be unclear in three cluster-randomised cross-over studies (Climo 2013; Milstone 2013; Noto 2015); some or all of the clusters in these studies were within the same hospital, which could influence recruitment to a particular ICU according to the current bathing regime. In Bleasdale 2007, the clusters were geographically separate which reduced this risk of recruitment bias; we judged this study to be at low risk.

Baseline imbalances (cluster trials only)

We judged one study to have low risk of bias because characteristics were reported, and were comparable, for each cluster (Bleasdale 2007). Three studies did not report baseline characteristics for each cluster, or we noted some differences between characteristics, and judged these to have an unclear risk of bias (Climo 2013; Milstone 2013; Noto 2015).

Loss of clusters (cluster trials only)

One study reported a loss of clusters (one unit withdrew, and two units were withdrawn from analysis by the study investigators because of low compliance with the protocol), and we judged this to introduce high risk of bias (Climo 2013).

Incorrect analysis (cluster trials only)

All cluster-randomised cross-over studies used appropriate analysis to account for the study design, and we judged them to have a low risk of bias for this domain (Bleasdale 2007; Climo 2013; Milstone 2013; Noto 2015).

Effects of interventions

See: [Summary of findings for the main comparison Bathing of the critically ill with chlorhexidine versus bathing with soap and water or non-antimicrobial washcloths for the prevention of hospital acquired infections](#)

We found data from eight studies, with a total of 24,472 participants, that compared bathing with a solution of chlorhexidine versus bathing with a solution of soap and water or non-antimicrobial washcloths. Study authors measured data for our primary outcome (hospital-acquired infections) and our secondary outcomes (mortality, and length of stay). We contacted study authors to provide clarification on missing data, and we included these data in the analysis where appropriate.

Chlorhexidine bathing versus bathing with soap and water or non-antimicrobial washcloths (seven studies; 24,023 participants)

Primary outcome: hospital-acquired infections

All studies collected and reported hospital-acquired infections during the intensive care unit (ICU) stay, and in the analysis we used data for bloodstream infections (BSI) (Bleasdale 2007; Climo 2013); hospital-acquired infections (Camus 2005); central line-associated bloodstream infections (CLABSI) (Milstone 2013); and composite infections of ventilator-associated pneumonia (VAP), CLABSI, and catheter-associated urinary tract infection (CAUTI) (Boonyasiri 2016); CLABSI, CAUTI, VAP, and clostridium difficile (Noto 2015); composite infections of BSI, CLABSI, urinary tract infection (UTI), CAUTI, and VAP (Pallotto 2018); and CAUTI, VAP, surgical site infection (SSI) and BSI (Swan 2016). Rates of bacteraemia were also reported in Milstone 2013; we did not include these data in analysis.

Details of the rate data, event data, and analysis process for this outcome are included in [Appendix 7](#) and [Appendix 8](#).

Despite a rate difference which indicated fewer hospital-acquired infections with chlorhexidine use, we are unsure whether using chlorhexidine for bathing critically ill people reduces hospital-acquired infections because the certainty of the evidence is very low (rate difference 1.70, 95% confidence interval (CI) 0.12 to 3.29; 21, 924 participants). See [Analysis 1.1](#).

We noted that one small study had a large influence on the rate difference for this outcome (Bleasdale 2007). We explored this in a sensitivity analysis, and we also explored the effect of using alternative design effects for one cluster-randomised cross-over study (Milstone 2013). See 'sensitivity analysis' below. Because of the results of the sensitivity analysis, we used the GRADE approach to downgrade the certainty of the evidence by one level for inconsistency. Most studies had a high risk of performance bias because personnel were aware of which product they were using to bathe participants, and we were concerned by other high risks of bias in individual studies; we downgraded by one level for study limitations. Participants in some studies may have had infections before randomisation; we downgraded by one level for indirectness. See [Summary of findings for the main comparison](#).

Secondary outcome: mortality

Six studies collected and reported data for mortality (Boonyasiri 2016; Camus 2005; Milstone 2013; Noto 2015; Pallotto 2018; Swan 2016). One study had excluded participants who died within 48 hours of randomisation (Boonyasiri 2016); we included these participants in the mortality data. Time points for data collection in other studies were: in-hospital mortality (Noto 2015; Swan 2016), and in-ICU mortality (Camus 2005; Pallotto 2018). The remaining study did not report a time point for data collection (Milstone 2013).

We analysed data for RCTs and cluster-randomised cross-over studies with generic inverse variance, and used standard errors imputed using an estimated design effect for two cluster-randomised cross-over studies (Milstone 2013; Noto 2015). We reported event data and details of the analysis process for these studies in [Appendix 9](#).

It is not clear whether using chlorhexidine for bathing critically ill people reduces mortality because the certainty of the evidence is

very low (OR 0.87, 95% CI 0.76 to 0.99; 15,798 participants). See [Analysis 1.2](#).

In a sensitivity analysis, we explored the effect of analysis decisions for the inclusion of two cluster-randomised cross-over studies ([Milstone 2013](#); [Noto 2015](#)). Consequently, we believe that the effect for mortality should be interpreted cautiously, and we used the GRADE approach to downgrade the certainty of the evidence by one level for inconsistency. Most studies had a high risk of performance bias because personnel were aware of which product they were using to bathe participants, and we were concerned by other high risks of bias in individual studies; we downgraded by one level for study limitations. Participants in some studies may have had infections before randomisation; we downgraded by one level for indirectness. See [Summary of findings for the main comparison](#).

Secondary outcome: length of stay in the intensive care unit

Six studies collected and reported length of stay in the ICU ([Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Noto 2015](#); [Pallotto 2018](#); [Swan 2016](#)).

We noted from visual inspection of the data, that reported ranges, SDs, and CIs in these studies were skewed; we decided it was not appropriate to combine data in analysis because of this. Individual study data are reported in [Table 1](#). We noted no evidence of any difference in length of stay in the ICU according to whether participants were bathed with chlorhexidine or soap and water.

It is unclear whether using chlorhexidine for bathing critically ill people reduces the length of stay in the ICU because the certainty of the evidence is very low.

We used the GRADE approach to downgrade the evidence by one level for imprecision because of skewed data reported by study authors. Most studies had a high risk of performance bias because personnel were aware of which product they were using to bathe participants, and we were concerned by other high risks of bias in individual studies; we downgraded by one level for study limitations. Participants in some studies may have had infections before randomisation; we downgraded by one level for indirectness. See [Summary of findings for the main comparison](#).

Secondary outcome: adverse effects

Seven studies reported participants who had skin irritation ([Bleasdale 2007](#); [Boonyasiri 2016](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Pallotto 2018](#); [Swan 2016](#)). Two studies reported adverse events of skin irritation but perceived these as not attributable to bathing ([Bleasdale 2007](#); [Climo 2013](#)); we have reported these data in [Characteristics of included studies](#).

It was not possible to combine data in meta-analysis for the remaining four studies. One study reported five participants with a mild skin reaction attributable to chlorhexidine ([Boonyasiri 2016](#)), but did not report whether data were collected for the control group. Another study reported 12 participants with a skin reaction attributable to chlorhexidine ([Milstone 2013](#)); skin reactions for the control group in this study were not reported according to whether they were attributable to the control. In [Pallotto 2018](#), one participant who was bathed with chlorhexidine had a mild skin reaction and chlorhexidine was discontinued in this participant.

In [Camus 2005](#), six participants in the control group had a skin reaction and six participants who had used chlorhexidine had a skin reaction. However, study authors had not reported how many of these were in the chlorhexidine with mupirocin group, which we had used as the intervention in the review. In [Swan 2016](#), there were 21 participants who were bathed with chlorhexidine, and 23 participants in the control group, who had skin reactions that were perceived as possibly or probably related to bathing.

We used the GRADE approach to downgrade the certainty of evidence for adverse events to very low. We downgraded by one level for study limitations; we judged some studies to have a high risk of performance bias, and some studies had high risks of other bias. Participants in some studies may have had infections before randomisation; we downgraded by one level for indirectness. We found few adverse events and we downgraded by one level for imprecision. See [Summary of findings for the main comparison](#).

Sensitivity analysis

1. Risk of bias

We assessed five of the eight studies included in our primary outcome to have unclear or high risk of selection bias ([Bleasdale 2007](#); [Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)).

Hospital-acquired infection

Analysis using only the remaining three parallel design studies showed little or no difference in infections according to bathing regime (rate difference 5.12, 95% CI -3.83 to 14.06).

Mortality

Analysis using only the remaining three parallel design studies did not alter interpretation of the effect.

2. Missing outcome data

All study authors reported losses and provided reasons. Four studies had reported the data as intention-to-treat and we had used these data in our meta-analyses ([Camus 2005](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)). The remaining studies had reported data only for those who were not lost during follow-up and we removed these studies from each of our meta-analyses. This did not alter direction or interpretation of the results.

3. Effects model for meta-analysis

When all available studies were included in the primary analysis, the conclusions remained the same regardless of whether a fixed-effect or random-effects model was used in meta-analysis.

4. Study design

Hospital-acquired infection

We included one cluster-randomised cross-over trial with only two clusters in the primary analysis ([Bleasdale 2007](#)). In our sensitivity analysis, we removed this study and found that the rate difference was reduced to indicate little or no difference in infections according to bathing regime (rate difference 1.26, 95% CI -0.21 to 2.72).

We included one study in which we imputed a SE using an estimated design effect ([Milstone 2013](#)). In our sensitivity analysis, we re-analysed the data by imputing an extremely conservative design effect of 5.06 (obtained by ignoring the cross-over effect and using

the formula $DE = 1 + (M - 1) ICC$ with $ICC = 0.05$ and $M = 495$, where DE = design effect, M = mean cluster size, ICC = intracluster correlation coefficient). We found that the rate difference was reduced in the sensitivity analysis, to indicate little or no difference in infections according to bathing regime when an extremely conservative design effect was used (rate difference 1.97, 95% CI -0.06 to 4.00).

Mortality

We included two cluster-randomised cross-over studies and used SE imputed using an estimated design effect (Milstone 2013; Noto 2015). In the sensitivity analysis, we re-analysed the data by imputing an extremely conservative design effect. We found little or no difference in mortality between different bathing regimes when extreme conservative design effects were used (OR 0.87, 95% CI 0.69 to 1.20).

DISCUSSION

Summary of main results

We identified eight studies: four randomised controlled trials (RCTs), which included 1537 randomised participants; and four cluster-randomised cross-over studies, which included 23 randomised intensive care units (ICUs) with a total of 22,935 participants. We identified one study awaiting classification, which was listed as completed in a clinical trial register but was not published.

Eight studies reported data for participants who had a hospital-acquired infection during their stay in the intensive care unit (ICU). Although the effect estimate showed fewer hospital-acquired infections with chlorhexidine bathing of critically ill people, the certainty of the evidence is very low. Six studies reported mortality (in hospital, in the ICU, and at 48 hours). Although the effect estimate showed reduced mortality with chlorhexidine bathing of critically ill people, the certainty of the evidence is very low. Six studies reported length of stay in the ICU. We noted that individual studies found no evidence of a difference in length of stay, and we did not conduct meta-analysis because data were skewed. We are uncertain whether using chlorhexidine for bathing of critically ill people reduced length of stay in the ICU because the certainty of the evidence is very low. Seven studies reported skin reactions as an adverse event, and five of these reported skin reactions which were thought to be attributable to the bathing solution. In these studies, data for skin irritation were reported inconsistently and we were unable to conduct meta-analysis; we are uncertain whether using chlorhexidine for bathing of critically ill people reduced adverse events, because the certainty of the evidence is very low. No other adverse events were reported in studies.

Overall completeness and applicability of evidence

We conducted a thorough search, including forward citation tracking of included studies, backward citation tracking of relevant reviews, and searches of grey literature. Included studies all compared chlorhexidine bathing with soap-and-water bathing or bathing with non-microbial washcloths, and included participants who were critically ill. We noted that participants in two studies had hospital-acquired infections before randomisation (Camus 2005; Swan 2016), and study authors in five studies did not report whether participants had hospital-acquired infections at baseline (Bleasdale 2007; Climo 2013; Milstone 2013; Noto 2015; Pallotto 2018). We believe that this introduces indirectness, and reduces the

applicability of the evidence for this review. Studies were published from 2005 to 2018, with five studies based in the USA, one in France, one in Italy, and one in Thailand.

Quality of the evidence

We used the GRADE approach to judge the evidence for each outcome to be of very low quality.

We considered study limitations identified during the 'Risk of bias' assessment. We noted some inconsistency in reporting between studies such that it was not possible to effectively judge all domains for each study. It was feasible to design a study so that personnel could be masked to the treatment allocation, yet only one study had effectively blinded personnel to the intervention and control, leading to a high risk of performance of bias across studies. Six studies did, however, make an effort to blind outcome assessors. We noted high participant attrition in some studies. We judged four studies with a cluster-randomised design to have an unclear risk of bias; these studies may have had differences at the unit-level of randomisation (i.e. between randomised ICUs). We noted differences in one study in which the chlorhexidine group received an additional treatment. We downgraded our assessment of the quality of the evidence, due to study limitations.

We noted some inconsistencies between results, and we used sensitivity analyses to explore this. We found that one small study had a large influence on the effect for hospital-acquired infections, and we found that decisions taken when estimating a design effect for some cluster-randomised cross-over studies may also have influenced results; therefore, we downgraded the evidence due to inconsistency. We also noted imprecision in individual study data for length of stay in the ICU, which were skewed. We were unable to explore potential differences between study participants (for example differences in illness severity or differences between adult and paediatric participants) because we had insufficient studies to conduct subgroup analysis. Whilst study participants were mostly applicable to our review question, we noted some indirectness because participants in some studies may have had hospital-acquired infections before randomisation. Because studies reported few adverse events, we downgraded our assessment of the evidence for this outcome because of imprecision. We were unable to assess the risk of publication bias because of lack of available data for this review.

Potential biases in the review process

We included four cluster-randomised cross-over studies in this review. We believe that this is an appropriate design for the study of infection practices. However, we did not anticipate this study design during preparation of the protocol, and so the methods used to analyse data from these studies were decided post-hoc. We used data reported by study authors if they were appropriately adjusted for both the clustering effect and the cross-over design; when we used estimation methods to calculate a design effect for the cluster studies, we assessed these decisions in sensitivity analysis.

We also used sensitivity analysis to explore the decisions to use a random-effects model for meta-analysis, and to use data reported only for participants who were not lost to follow-up in three of our included studies, neither of which influenced interpretation of our results. There were insufficient studies to explore risks of selection bias in our analyses. We did not attempt to consider other

factors that may have impacted on our data. The decision of which treatment and control group to include in one study meant that participants in the chlorhexidine group for this study were also treated with mupirocin, which the control group did not receive (Camus 2005); this may have acted as a confounder for these data, which we did not explore. We included three studies in which it was noted that some participants had a hospital-acquired infection and we did not assess whether this influenced our results, nor did we explore the impact of two large multi-centre studies on our data (Milstone 2013; Noto 2015).

We conducted the review according to the protocol, with two reviewers independently assessing studies for eligibility, extracting data and carrying out the 'Risk of bias' assessment.

Agreements and disagreements with other studies or reviews

There have been several systematic reviews that have assessed the effect of chlorhexidine bathing on the critically ill. Reviews have previously concluded that chlorhexidine bathing reduces risk of infection in the ICU (Chen 2015; Choi 2015; Huang 2016; Kim 2016; O'Horo 2012). These reviews have collected event data for specific infections (bloodstream infections, central line-associated bloodstream infections, ventilator-associated pneumonia, Methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococci*) rather than a composite outcome for the number of participants with any hospital-acquired infection.

Most notably, these systematic reviews include both RCTs and non-randomised study designs; and one review noted that this effect was not consistent when non-randomised studies were excluded from analysis (Chen 2015). A review that had only included RCTs, did not exclude Huang 2013; this study had a large sample size but did not compare chlorhexidine with soap and water. It is possible that the results of our analyses are dependent on our restriction to RCTs.

AUTHORS' CONCLUSIONS

Implications for practice

It is not clear whether bathing with chlorhexidine reduces hospital-acquired infections, mortality or length of stay in the intensive care unit, or whether chlorhexidine use results in more skin reactions, because the certainty of the evidence is very low. One study is awaiting classification and two studies are ongoing; we do not know if inclusion of these studies in future updates of this Cochrane Review will increase our certainty in the results of the review.

Implications for research

Additional research is needed to evaluate whether chlorhexidine bathing may reduce hospital-acquired infections in the intensive care unit. We recommend that studies are sufficiently powered and methodologically robust, and that attention is paid to reduce the risk of performance bias through blinding of personnel. Cluster-randomised studies and cross-over trials would benefit from reporting data in more detail, including important parameters such as the intracluster correlation coefficient and interperiod correlation. Some consensus on the reporting of hospital-acquired infection rates, for example through the adoption of a core outcome set for trials of infection prevention, would also be helpful.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of peer referees who commented on the protocol, the review or both: Zipporah Ihezor-Ejiofor, Julie Bruce, Sharon Van Wicklin, Mahmoud Alkhatib, Amanda Roberts and Gill Norman and to thank Megan Pricor and Jessica Sharp who copy-edited the protocol and the review respectively. The authors would also like to thank Phil Alderson, Andrew Butler and David Evans for their work on the protocol.

REFERENCES

References to studies included in this review

Bleasdale 2007 {published data only}

Bleasdale SC, Trick WE, Gonzalez IM, Lyles RD, Hayden MK, Weinstein RA. Effectiveness of chlorhexidine bathing to reduce catheter-associated bloodstream infections in medical intensive care unit patients. *Archives of Internal Medicine* 2007;**167**(19):2073-9. [PUBMED: 17954801]

Boonyasiri 2016 {published data only}

Boonyasiri A, Thaisiam P, Permpikul C, Judaeng T, Suiwongsa B, Apiradeewajeset N, et al. Effectiveness of chlorhexidine wipes for the prevention of multidrug-resistant bacterial colonization and hospital-acquired infections in intensive care unit patients: a randomized trial in Thailand. *Infection Control and Hospital Epidemiology* 2016;**37**(3):245-53. [PUBMED: 26894621]

Camus 2005 {published data only}

* Camus C, Bellissant E, Seville V, Perrotin D, Garo B, Legras A, et al. Prevention of acquired infections in intubated patients with the combination of two decontamination regimens. *Critical Care Medicine* 2005;**33**(2):307-14. [PUBMED: 15699832]

Camus C, Seville V, Legras A, Garo B, Renault A, Le Corre P, et al. Mupirocin/chlorhexidine to prevent methicillin-resistant *Staphylococcus aureus* infections: post hoc analysis of a placebo-controlled, randomized trial using mupirocin/chlorhexidine and polymyxin/tobramycin for the prevention of acquired infections in intubated patients. *Infection* 2014;**42**(3):493-502. [PUBMED: 24464791]

Climo 2013 {published data only}

Climo MW, Yokoe DS, Warren DK, Perl TM, Bolon M, Herwaldt LA, et al. Effect of daily chlorhexidine bathing on hospital-acquired infection. *New England Journal of Medicine* 2013;**368**(6):533-42. [PUBMED: 23388005]

Milestone 2013 {published data only}

Milestone AM, Elward A, Song X, Zerr DM, Orscheln R, Speck K, et al. Daily chlorhexidine bathing to reduce bacteraemia in critically ill children: a multicentre, cluster-randomised, crossover trial. *Lancet* 2013;**381**(9872):1099-1106. [PUBMED: 23363666]

Noto 2015 {published data only}

Noto M, Domenico H, Talbot T, Byrne D, Wheeler A. Healthcare-associated infections and chlorhexidine bathing - a pragmatic cluster-randomized trial. *Critical Care Medicine* 2014;**42**(12):A1478-9. [Abstract 493]

* Noto MJ, Domenico HJ, Byrne DW, Talbot T, Rice TW, Bernard GR, et al. Chlorhexidine bathing and health care-associated infections: a randomized clinical trial. *Journal of the American Medical Association* 2015;**313**(4):369-78. [PUBMED: 25602496]

Pallotto 2018 {published data only}

Pallotto C, Fiorio M, De Angelis V, Ripoli A, Franciosini E, Quondam Girolamo L, et al. Daily bathing with 4% chlorhexidine gluconate in intensive care settings: a randomized controlled

trial. *Clinical Microbiology and Infection* 2018 Sept 26 [Epub ahead of print]. [DOI: [10.1016/j.cmi.2018.09.012](https://doi.org/10.1016/j.cmi.2018.09.012); PUBMED: 30267930]

Swan 2016 {published data only}

Bui L, Badawi N, Swan J, Bersamin J. Preliminary results of a randomized controlled trial comparing the incidence of nosocomial infections with chlorhexidine bathing versus standard bathing in the surgical intensive care unit. *Journal of the American Pharmacists Association* 2013;**53**(2):e82-e83.

NCT01640925. Randomized controlled trial of 2% chlorhexidine bathing on nosocomial infections in the surgical intensive care unit. clinicaltrials.gov/show/NCT01640925 (first received 16 July 2012).

Swan J, Bui L, Pham V, Shirkey B, Graviss E, Hai S, et al. RCT of chlorhexidine vs. soap & water bathing for prevention of hospital-acquired infections in SICU. *Critical Care Medicine* 2014;**42**(12 (Suppl 1)):A1369-70.

* Swan JT, Ashton CM, Bui LN, Pham VP, Shirkey BA, Blackshear JE, et al. Effect of chlorhexidine bathing every other day on prevention of hospital-acquired infections in the surgical ICU: a single-center, randomized controlled trial. *Critical Care Medicine* 2016;**44**(10):1822-32. [PUBMED: 27428384]

References to studies excluded from this review

Camus 2011 {published data only}

Camus C, Bellissant E, Legras A, Renault A, Gacouin A, Lavoué S, et al. Randomized comparison of 2 protocols to prevent acquisition of methicillin-resistant *Staphylococcus aureus*: results of a 2-center study involving 500 patients. *Infection Control and Hospital Epidemiology* 2011;**32**(11):1064-72. [PUBMED: 22011532]

Choi 2012 {published data only}

Choi JS, Yeon JH. Effects of perineal care in preventing catheter associated urinary tract infections (CAUTI) in intensive care units (ICU). *Journal of Korean Academy of Fundamentals of Nursing* 2012;**19**(2):223-32.

Cunha 2008 {published data only}

Cunha ML, Procianoy RS, Franceschini DT, Oliveira LL, Ballin A, Livshiz V, et al. Effect of the first bath with chlorhexidine on skin colonization with *Staphylococcus aureus* in normal healthy term newborns. *Scandinavian Journal of Infectious Diseases* 2008;**40**(8):615-20. [PUBMED: 18979599]

Dean 2011 {published data only}

Dean R, Dillworth J, Phillips M. Assessment of daily bathing protocols: a comparison of chlorhexidine solution and chlorhexidine impregnated cloths. *Critical Care Medicine* 2011;**39**(12 (Suppl)):142.

Derde 2014 {published data only}

Derde LP, Cooper BS, Brun-Buisson C, Bonten MJ. Reducing acquisition of resistant bacteria in intensive cares: a European

cluster randomised trial. *Clinical Microbiology and Infection* 2012;**18**(Suppl 3):712.

* Derde LP, Cooper BS, Goossens H, Malhotra-Kumar S, Willems RJ, Gniadkowski M, et al. Interventions to reduce colonisation and transmission of antimicrobial-resistant bacteria in intensive care units: an interrupted time series study and cluster randomised trial. *Lancet Infectious Diseases* 2014;**14**(1):31-9. [PUBMED: 24161233]

Derde LP, Dautzenberg MJ, Van Duijn PJ, Brun-Buisson C, Bonten MJ. Antimicrobial resistance in ICU-acquired bacteraemias in 13 European intensive care units. *Clinical Microbiology and Infection* 2011;**17**(Suppl 4):S423. [Poster 1490]

Duzkaya 2017 {published data only}

Duzkaya DS, Uysal G, Bozkurt G, Yakut T, Citak A. Povidone-iodine, 0.05% chlorhexidine gluconate, or water for periurethral cleaning before indwelling urinary catheterization in a pediatric intensive care: a randomized controlled trial. *Journal of Wound, Ostomy, and Continence Nursing* 2017;**44**(1):84-8. [PUBMED: 27824737]

Huang 2013 {published data only}

Hayden MK, Lolans K, Haffenreffer K, Avery TR, Kleinman K, Li H, et al. Chlorhexidine and mupirocin susceptibility of methicillin-resistant staphylococcus aureus isolates in the REDUCE-MRSA trial. *Journal of Clinical Microbiology* 2016;**54**(11):2735-42. [PUBMED: 27558180]

* Huang SS, Septimus E, Kleinman K, Moody J, Hickok J, Avery TR, et al. Targeted versus universal decolonization to prevent ICU infection. *The New England Journal of Medicine* 2013;**368**(24):2255-65. [PUBMED: 23718152]

Lowe 2017 {published data only}

Lowe CF, Lloyd-Smith E, Sidhu B, Ritchie G, Sharma A, Jang W, et al. Reduction in hospital-associated methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant *Enterococcus* with daily chlorhexidine gluconate bathing for medical inpatients. *American Journal of Infection Control* 2017;**45**(3):255-9. [PUBMED: 27938986]

Sankar 2009 {published data only}

Sankar MJ, Paul VK, Kapil A, Kalaivani M, Agarwal R, Darmstadt GL, et al. Does skin cleansing with chlorhexidine affect skin condition, temperature and colonization in hospitalized preterm low birth weight infants? A randomized clinical trial. *Journal of Perinatology* 2009;**29**(12):795-801. [PUBMED: 19710679]

References to studies awaiting assessment

ChiCTR-TRC-13004164 {published data only}

ChiCTR-TRC-13004164. The efficacy of daily chlorhexidine body bathing for reducing nosocomial infections in intensive care units. www.chictr.org.cn/showproj.aspx?proj=5404 (first received 13 December 2013).

References to ongoing studies

IRCT2017030932293N1 {published data only}

IRCT2017030932293N1. Chlorhexidine effect on skin colonization [The effect of daily 2% chlorhexidine bathing, on colonization of the skin of patients admitted to Khatam's hospital ICU of the Zahedan University of Medical Sciences in 2017]. apps.who.int/trialsearch/Trial2.aspx?TrialID=IRCT2017030932293N1 (first received 23 March 2017).

NCT02870062 {published data only}

NCT02870062. Impact of daily bathing with chlorhexidine in the critical patient [Impact of daily bathing with chlorhexidine in the critical patient: colonization and environment]. clinicaltrials.gov/ct2/show/NCT02870062 (first received 7 August 2016).

Additional references

Afonso 2013

Afonso E, Llauradó M, Gallart E. The value of chlorhexidine gluconate wipes and prepacked washcloths to prevent the spread of pathogens: a systematic review. *Australian Critical Care* 2013;**26**(4):158-66. [PUBMED: 23827390]

Afonso 2016

Afonso E, Blot K, Blot S. Prevention of hospital-acquired bloodstream infections through chlorhexidine gluconate-impregnated washcloth bathing in intensive care units: a systematic review and meta-analysis of randomised crossover trials. *Euro Surveillance* 2016;**21**(46):30400. [PUBMED: 27918269]

Bereket 2012

Bereket W, Hemalatha K, Getenet B, Wondwossen T, Solomon A, Zeynudin A, et al. Update on bacterial nosocomial infections. *European Review for Medical and Pharmacological Sciences* 2012;**16**(8):1039-44. [PUBMED: 22913154]

Calogiuri 2013

Calogiuri GF, Di Leo E, Trautmann A, Nettis E, Ferrannini A, Vacca A. Chlorhexidine hypersensitivity: a critical and updated review. *The Journal of Allergy and Therapy* 2013;**4**(141):2.

Chen 2015

Chen W, Cao Q, Li S, Li H, Zhang W. Impact of daily bathing with chlorhexidine gluconate on ventilator associated pneumonia in intensive care units: a meta-analysis. *Journal of Thoracic Disease* 2015;**7**(4):746-53. [PUBMED: 25973242]

Choi 2015

Choi EY, Park D-A, Kim HJ, Park J. Efficacy of chlorhexidine bathing for reducing healthcare associated bloodstream infections: a meta-analysis. *Annals of Intensive Care* 2015;**5**(1):31. [PUBMED: 26445950]

Covidence 2017 [Computer program]

Veritas Health Innovation. Covidence. Version accessed 18 October 2017. Melbourne, Australia: Veritas Health Innovation, 2017.

Deeks 2017

Deeks JJ, Higgins JP, Altman DG, editor(s) on behalf of the CSMG. Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins JP, Churchill R, Chandler J, Cumpston MS editor(s), Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0 (updated June 2017), The Cochrane Collaboration, 2017. Available from www.training.cochrane.org/handbook.

Derde 2012

Derde LP, Dautzenberg MJ, Bonten MJ. Chlorhexidine body washing to control antimicrobial-resistant bacteria in intensive care units: a systematic review. *Intensive Care Medicine* 2012;**38**(6):931-9. [PUBMED: 22527065]

Egger 1997

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**(7109):629-34. [PUBMED: 9310563]

Endnote 2011 [Computer program]

Thomson Reuters. Endnote. Version X5. New York: Thomson Reuters, 2011.

Faria 2007

Faria S, Sodano L, Gjata A, Dauri M, Sabato A, Bilaj A, et al. The first prevalence survey of nosocomial infections in the University Hospital Centre 'Mother Teresa' of Tirana, Albania. *Journal of Hospital Infection* 2007; Vol. 65, issue 3:244-50.

Frost 2016

Frost SA, Alogso MC, Metcalfe L, Lynch JM, Hunt L, Sanghavi R, et al. Chlorhexidine bathing and health care-associated infections among adult intensive care patients: a systematic review and meta-analysis. *Critical Care* 2016;**20**(1):379. [PUBMED: 27876075]

Genuit 2001

Genuit T, Bochicchio G, Napolitano LM, McCarter RJ, Roghman MC. Prophylactic chlorhexidine oral rinse decreases ventilator-associated pneumonia in surgical ICU patients. *Surgical Infections* 2001;**2**(1):5-18. [PUBMED: 12594876]

GRADE 2013

Schünemann H, Brożek J, Guyatt G, Oxman A, editor(s). Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Working Group, 2013. Available from gdt.guidelinedevelopment.org/app/handbook/handbook.html.

GRADEpro 2015 [Computer program]

McMaster University (developed by Evidence Prime). GRADEpro GDT. Version accessed 9 October 2018. Hamilton (ON): McMaster University (developed by Evidence Prime), 2015.

Guyatt 2008

Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schünemann HJ. What is 'quality of evidence' and why is it important to clinicians?. *BMJ* 2008;**336**(7651):995-8. [PUBMED: 18456631]

Guyatt 2011a

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence - imprecision. *Journal of Clinical Epidemiology* 2011;**64**(12):1283-93. [PUBMED: 21839614]

Guyatt 2011b

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence - inconsistency. *Journal of Clinical Epidemiology* 2011;**64**(12):1294-302. [PUBMED: 21803546]

Health Protection Agency 2016

Health Protection Agency. Healthcare associated infections (HAI): point prevalence survey, England. Updated August 2016. www.gov.uk/government/publications/healthcare-associated-infections-hcai-point-prevalence-survey-england (accessed 24 September 2018).

Hibbard 2005

Hibbard JS. Analyses comparing the antimicrobial activity and safety of current antiseptic agents: a review. *Journal of Infusion Nursing* 2005;**28**(3):194-207. [PUBMED: 15912075]

Higgins 2011

Higgins JP, Green S, editor(s). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from handbook.cochrane.org.

Higgins 2017

Higgins JP, Altman DG, Sterne JA, editor(s). Chapter 8: Assessing risk of bias in included studies. In: Higgins JP, Churchill R, Chandler J, Cumpston MS, editor(s), Cochrane Handbook for Systematic Reviews of Interventions version 5.2.0 (updated June 2017), The Cochrane Collaboration, 2017. Available from www.training.cochrane.org/handbook.

Horner 2012

Horner C, Mawer D, Wilcox M. Reduced susceptibility to chlorhexidine in staphylococci: is it increasing and does it matter?. *Journal of Antimicrobial Chemotherapy* 2012;**67**(11):2547-59. [PUBMED: 22833635]

Huang 2016

Huang HP, Chen B, Wang HY, He M. The efficacy of daily chlorhexidine bathing for preventing healthcare-associated infections in adult intensive care units. *The Korean Journal of Internal Medicine* 2016;**31**:1159-70. [PUBMED: 27048258]

Ibrahim 2001

Ibrahim EH, Tracy L, Hill C, Fraser VJ, Kollef MH. The occurrence of ventilator-associated pneumonia in a community hospital: risk factors and clinical outcomes. *Chest* 2001;**120**(2):555-61. [PUBMED: 11502658]

Ider 2010

Ider B-E, Clements A, Adams J, Whitby M, Muugolog T. Prevalence of hospital-acquired infections and antibiotic use in two tertiary Mongolian hospitals. *Journal of Hospital Infection* 2010; Vol. 75, issue 3:214-9.

Inweregbu 2005

Inweregbu K, Dave J, Pittard A. Nosocomial infections. *Continuing Education in Anaesthesia, Critical Care and Pain* 2005;**5**(1):14-7.

Kelly 2012

Kelly KN, Monson JR. Hospital-acquired infections. *Surgery* 2012;**30**(12):640-4.

Kim 2016

Kim HY, Lee WK, Na S, Roh YH, Shin CS, Kim J. The effects of chlorhexidine gluconate bathing on health care-associated infection in intensive care units: a meta-analysis. *Journal of Critical Care* 2016;**32**:126-37.

Lefebvre 2011

Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JP, Green S, editor(s). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from handbook.cochrane.org.

Liberati 2009

Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;**339**:b2700. [PUBMED: 19622552]

Magill 2014

Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, et al. Multistate point-prevalence survey of health care-associated infections. *New England Journal of Medicine* 2014;**370**(13):1198-208. [PUBMED: 24670166]

Mahjoub 2015

Mahjoub M, Bouafia N, Bannour W, Masmoudi T, Bouriga R, Hellali R, et al. Healthcare-associated infections in a Tunisian university hospital: from analysis to action. *Pan African Medical Journal* 2015;**20**:197. [PUBMED: 26113928]

McDonnell 1999

McDonnell G, Russell AD. Antiseptics and disinfectants: activity, action, and resistance. *Clinical Microbiology Reviews* 1999;**12**(1):147-79. [PUBMED: 9880479]

O'Horo 2012

O'Horo JC, Silva GL, Munoz-Price S, Safdar N. The efficacy of daily bathing with chlorhexidine for reducing healthcare-associated bloodstream infections: a meta-analysis. *Infection Control and Hospital Epidemiology* 2012;**33**(3):257-67. [PUBMED: 22314063]

Puig Silla 2008

Puig Silla M, Montiel Company JM, Almerich Silla JM. Use of chlorhexidine varnishes in preventing and treating periodontal disease: a review of the literature. *Medicina Oral, Patologia Oral y Cirugia Bucal* 2008;**13**(4):E257-60. [PUBMED: 18379452]

Review Manager 2014 [Computer program]

Nordic Cochrane Centre, The Cochrane Collaboration. Review Manager 5 (RevMan 5). Version 5.3. Copenhagen: Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Sacks 2014

Sacks GD, Diggs BS, Hadjizacharia P, Green D, Salim A, Malinoski DJ. Reducing the rate of catheter-associated bloodstream infections in a surgical intensive care unit using the Institute for Healthcare Improvement Central Line Bundle. *American Journal of Surgery* 2014;**207**(6):817-23. [PUBMED: 24576582]

Shah 2016

Shah HN, Schwartz JL, Luna G, Cullen DL. Bathing with 2% chlorhexidine gluconate: evidence and costs associated with central line-associated bloodstream infections. *Critical Care Nursing Quarterly* 2016;**39**(1):42-50. [PUBMED: 26633158]

Shi 2013

Shi Z, Xie H, Wang P, Zhang Q, Wu Y, Chen E, et al. Oral hygiene care for critically ill patients to prevent ventilator-associated pneumonia. *Cochrane Database of Systematic Reviews* 2013, Issue 8. [DOI: [10.1002/14651858.CD008367.pub2](https://doi.org/10.1002/14651858.CD008367.pub2)]

SIGN 2018

Scottish Intercollegiate Guidelines Network (SIGN). Search filters. www.sign.ac.uk/search-filters.html (accessed 24 September 2018).

Sinha 2015

Sinha A, Sazawal S, Pradhan A, Ramji S, Opiyo N. Chlorhexidine skin or cord care for prevention of mortality and infections in neonates. *Cochrane Database of Systematic Reviews* 2015, Issue 3. [DOI: [10.1002/14651858.CD007835.pub2](https://doi.org/10.1002/14651858.CD007835.pub2)]

Sterne 2017

Sterne JA, Egger M, Moher D, Boutron I, editor(s). Chapter 10: Addressing reporting biases. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS, editor(s), *Cochrane Handbook for Systematic Reviews of Interventions* version 5.2.0 (updated June 2017), The Cochrane Collaboration, 2017. Available from www.training.cochrane.org/handbook.

Vincent 2009

Vincent JL, Rello J, Marshall J, Silva E, Anzueto A, Martin CD, et al. International study of the prevalence and outcomes of infection in intensive care units. *Journal of the American Medical Association* 2009;**302**(21):2323-9. [PUBMED: 19952319]

Volk 2002

Volk HD. Immunodepression in the surgical patient and increased susceptibility to infection. *Critical Care* 2002;**6**(4):279-81. [PUBMED: 12225595]

Wade 1991

Wade JJ, Casewell MW. The evaluation of residual antimicrobial activity on hands and its clinical relevance. *The Journal of Hospital Infection* 1991;**18**(Suppl B):23-8. [PUBMED: 1679443]

WHO 2011

World Health Organization (WHO). Report on the burden of endemic health care-associated infection worldwide. 2011. apps.who.int/iris/bitstream/10665/80135/1/9789241501507_eng.pdf (accessed 24 September 2018).

WHO 2017

World Health Organization (WHO). WHO model lists of essential medicines. Amended August 2017. www.who.int/medicines/publications/essentialmedicines/en/ (accessed 24 September 2018).

References to other published versions of this review
Lewis 2016

Lewis SR, Butler AR, Evans DJ, Alderson P, Smith AF. Chlorhexidine bathing of the critically ill for the prevention of hospital-acquired infection. *Cochrane Database of Systematic Reviews* 2016, Issue 6. [DOI: [10.1002/14651858.CD012248](https://doi.org/10.1002/14651858.CD012248)]

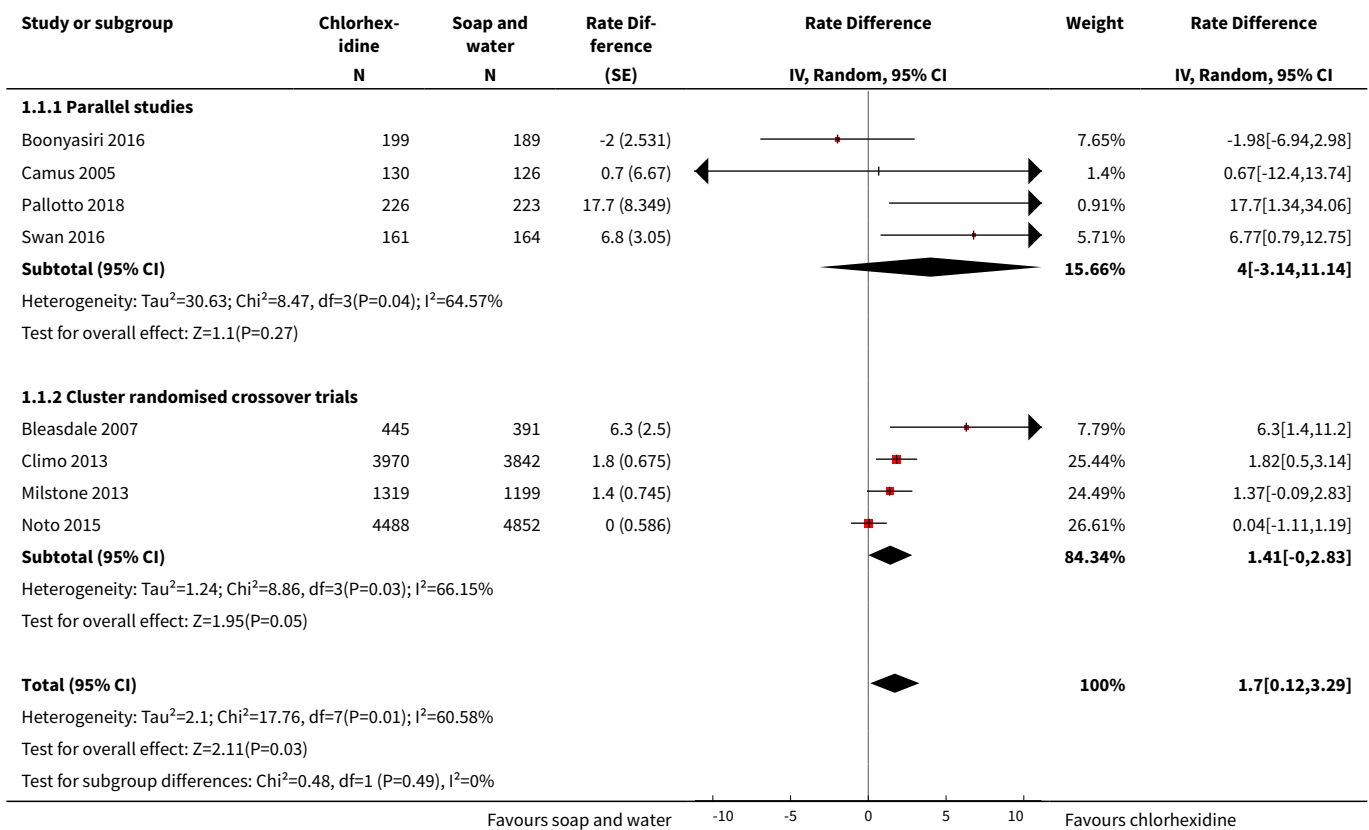
* Indicates the major publication for the study

CHARACTERISTICS OF STUDIES
Characteristics of included studies [ordered by study ID]
Bleasdale 2007

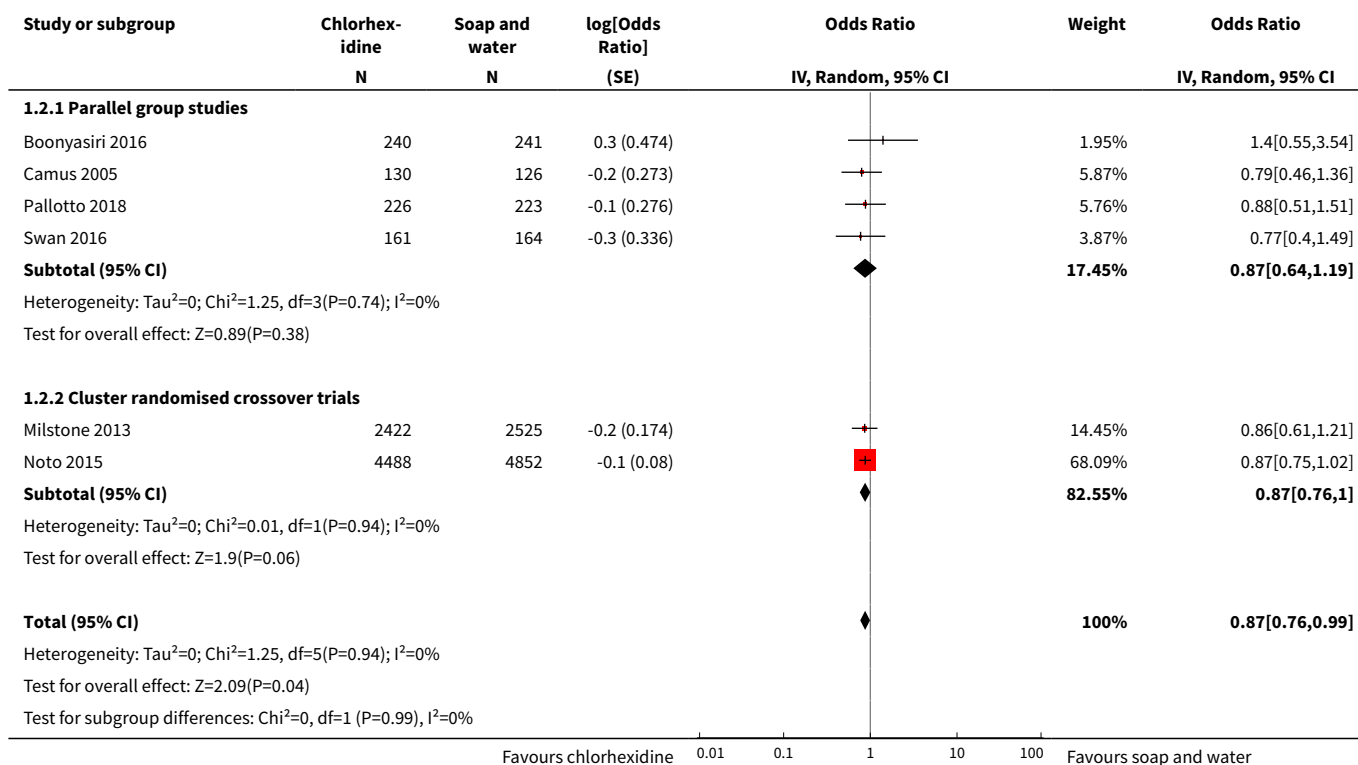
Methods	Cluster-randomised cross-over study; single centre Setting: 2 MICUs, USA Unit of randomisation: MICU 28-week initial phase followed by alternative bathing routine for 24 weeks 2-week washout period
Participants	<p>Total number of randomised participants: 836</p> <p>Inclusion criteria: all people who attended one of 2 MICUs</p> <p>Exclusion criteria: not reported</p> <p>Number of participants with an HAI before randomisation: not reported</p> <p>Baseline characteristics</p> <p>Chlorhexidine gluconate (n = 391; 3 excluded due to skin rash, but use of ITT, therefore number analysed = 391)</p> <ol style="list-style-type: none"> Age, mean (SD): 53 (± 16) years Gender, M/F: 234/157 APACHE II, mean (SD): 22.4 (± 7) <p>Soap and water (n = 445; no losses)</p> <ol style="list-style-type: none"> Age, mean (SD): 52 (± 15) years Gender M/F: 266 /179 APACHE II, mean (SD): 21.5 (± 7)
Interventions	Chlorhexidine gluconate (2%) <ol style="list-style-type: none"> Administration: 8 CHG-impregnated washcloths, daily cleaning everywhere except face, cloths warmed for participant comfort, 2 non-medicated cloths to clean the participants' faces Soap and water <ol style="list-style-type: none"> Administration: daily bathing in warm water, with 10 terry cloth washcloths and soap Concurrent decolonisation strategies: use of sterile catheter insertion policy (without CHG coating), full barrier drapes and insertion site disinfection with CHG

Outcome or subgroup title	No. of studies	No. of participants	Statistical method	Effect size
1.1 Parallel studies	4	1418	Rate Difference (Random, 95% CI)	4.00 [-3.14, 11.14]
1.2 Cluster randomised crossover trials	4	20506	Rate Difference (Random, 95% CI)	1.41 [-0.00, 2.83]
2 Mortality using adjusted data	6	15798	Odds Ratio (Random, 95% CI)	0.87 [0.76, 0.99]
2.1 Parallel group studies	4	1511	Odds Ratio (Random, 95% CI)	0.87 [0.64, 1.19]
2.2 Cluster randomised crossover trials	2	14287	Odds Ratio (Random, 95% CI)	0.87 [0.76, 1.00]

Analysis 1.1. Comparison 1 Chlorhexidine bathing versus soap-and-water bathing, Outcome 1 Hospital-acquired infection.



Analysis 1.2. Comparison 1 Chlorhexidine bathing versus soap-and-water bathing, Outcome 2 Mortality using adjusted data.



ADDITIONAL TABLES

Table 1. Data for length of stay

Study ID	Study design	Data reported by study authors	Chorhexidine group	Control group	Inference
Boonyasiri 2016	Parallel	Median (range) length of ICU stay, in days	9 (3 to 212)	10 (3 to 136)	P = 0.42
Camus 2005	Parallel	Median (range) length of ICU stay, in days	15 (3 to 132)	16 (3 to 83)	"Not significantly different"
Climo 2013	Cluster-randomised cross-over	Mean length of ICU stay, in days	6.4	6.4	P = 0.53 (unadjusted)
Noto 2015	Cluster-randomised cross-over	Mean (95% CI) length of ICU stay, in days	2.56 (1.24 to 5.09)	2.39 (1.21 to 4.95)	Difference (95% CI) = 0.169 (-0.01 to 0.321) (unadjusted using Mann-Whitney U)
Pallotto 2018	Parallel	Median (IQR) length of ICU stay, in days	4 (2 to 8)	4 (2 to 7)	P > 0.05

Warren DK, Prager M, Munigala S, Wallace MA, Kennedy CR, Bommarito KM, et al. Prevalence of qacA/B genes and mupirocin resistance among methicillin-resistant staphylococcus aureus (MRSA) isolates in the setting of chlorhexidine bathing without mupirocin. *Infection Control and Hospital Epidemiology* 2016;37(5):590-7

Appendix 7. Statistical analysis details: hospital-acquired infections

Study ID	Study design	Analysis reported by study author	Definition of outcome used in the review for 'hospital acquired infection'	Data reported by study authors	Manipulation by review authors
Bleasdale 2007	Cluster-randomised cross-over trial with 2 clusters and 836 participants	Multivariate models that include a term for geographical unit	BSI	Rate (per 1000 patient days) 10.4 vs 4.1, 95% CI for rate difference 1.2 to 11	SE for rate difference calculated from CI
Boonyasiri 2016	Parallel group trial of 481 participants	Individual incidence	VAP, CLABSI, and CAUTI	28 infections during 3284 patient days vs 29 infections during 2759 patient days (adding up infections and using mean ICU stay to calculate patient days)	Rate difference and associated SE calculated using section 9.4.8 Higgins 2011
Camus 2005	2x2 factorial trial with 256 participants in two relevant arms	Number of infections and patient days	Acquired infections	87 infections during 1961 patient days vs 87 infections during 1991 patient days	Rate difference and associated SE calculated using section 9.4.8 Higgins 2011
Climo 2013	Cluster-randomised cross-over trial with 9 clusters and 7727 participants	GEE (according to SAP)	Hospital acquired BSI	Rate (per 1000 patient days) 6.60 vs 4.78, P = 0.007	P value used to calculate Z-value, and rate difference/ Z gives estimate of SE
Milestone 2013	Cluster-randomised cross-over trial with 10 clusters and 4947 participants	Poisson regression adjusted for cluster and time	CLABSI	Rate (per 1000 patient days) 3 vs 1.63, rate ratio 0.52 (95% CI 0.26 to 1.08), adjusted CI 0.25 to 1.08	Unadjusted and adjusted CI used to calculate SE for log rate ratio: from these design effect = 1.03 ² applied to inflate SE of rate difference
Noto 2015	Cluster-randomised cross-over trial with 5 clusters and 9340 participants	Supplementary materials present group level analysis of clusters	Composite outcome including CLABSI, CAUTI, VAP and Clostridium difficile	Rate (per 1000 patient days) 3.35 vs 3.31, 95% CI for rate difference (-1.19 to 1.11)	SE for rate difference calculated from CI
Palotto 2018	Parallel group trial of 449 participants	Number of infections per patient days	Composite outcome including BSI, CLABSI,	Rate (per 1000 patient days) 40.9 vs 23.2, P = 0.034	P value used to calculate Z-value, and rate difference/ Z gives estimate of SE

(Continued)

			UTI, CAUTI, and VAP		
Swan 2016	Parallel group trial of 350 participants	Hazard ratios and risk difference	Composite outcome of CAUTI, VAP, SSI, BSI	35 infections during 2416 patient days vs 18 infections during 2332 patient days (supplementary digital content)	Rate difference and associated SE calculated using section 9.4.8 Higgins 2011

BSI: blood stream infection; CAUTI: catheter associated urinary tract infection; CI: confidence interval; CLABSI: central line associated blood stream infection; GEE: general estimating equation; ICU: intensive care unit; SAP: statistical analysis plan; SE: standard error; UTI: urinary tract infection; VAP: ventilator-acquired pneumonia

Appendix 8. Analysis of rate differences: hospital-acquired infections

Study ID	Rate difference (control – treatment)	95% CI	SE
Bleasdale 2007	6.3	1.2 to 11	2.5
Boonyasiri 2016	-1.984		2.53
Camus 2005	0.668		6.68
Climo 2013	1.82	0.43 to 3.13	0.675
Milstone 2013	1.37	-0.24 to 2.25	0.756
Noto 2015	0.04	-1.11 to 1.19	0.586
Pallotto 2018	17.7	1.34 to 34.06	8.349
Swan 2016	6.768		3.050

CI: confidence interval; SE: standard error

Appendix 9. Statistical analysis details: mortality

Study ID	Data reported by study authors	Manipulation by review authors
Milstone 2013	88/2525 vs 73/2422 with unadjusted RD -0.48(95% CI -0.147 to 0.51)	Design effect = 1.03 ² calculated for primary outcome. This was applied to inflate standard error of log OR
Noto 2015	449/4852 vs 367/4488 with unadjusted CII for RD -1.07(95% CI -2.22 to 0.07)	Adjusted and unadjusted CI presented for RD for primary outcome: these were used to estimate design effect = 1.09 ² which was applied to inflate SE of log OR

CI: confidence interval; OR: odds ratio; RD: risk difference; SE: standard error

HISTORY

Protocol first published: Issue 6, 2016

Review first published: Issue 8, 2019

Date	Event	Description
30 March 2020	Amended	Minor edit to 'Summary of Findings' table.

CONTRIBUTIONS OF AUTHORS

Sharon R Lewis: conceived, designed and coordinated the review; extracted data; checked the quality of data extraction; analysed or interpreted data; undertook and checked quality assessment; produced the first draft of the review; contributed to writing and editing the review; and wrote to study authors/experts/companies.

Oliver J Schofield-Robinson: extracted data; checked the quality of data extraction; analysed or interpreted data; undertook and checked quality assessment; produced the first draft of the review; and wrote to study authors/experts/companies.

Sarah Rhodes: analysed or interpreted data; performed statistical analysis; checked the quality of the statistical analysis; produced the first draft of the review; and contributed to writing or editing the review.

Andrew Smith: checked the quality of data extraction; checked quality assessment; produced the first draft of the review; advised on the review; secured funding; approved the final review prior to submission; and is a guarantor of the review.

Contributions of the editorial base

Joan Webster (Editor): edited the protocol; advised on methodology, interpretation and content; approved the final protocol prior to submission.

Jo Dumville (Coordinating Editor): edited the review; advised on methodology, interpretation and content; approved the final review prior to submission.

Gill Rizzello (Managing Editor): coordinated the editorial process; advised on content; edited the protocol and the review.

Reetu Child, Naomi Shaw and Sophie Bishop (Information Specialists): designed the search strategy, ran the search and edited the search methods section.

Ursula Gonthier (Editorial Assistant): edited the Plain language summary and reference sections of the review.

DECLARATIONS OF INTEREST

Sharon R Lewis: my work on this review is funded by the National Institute for Health Research (NIHR) Cochrane programme grant 13/89/16 ('Back to normal': speed and quality of recovery after surgery, major injury and critical care).

Oliver J Schofield-Robinson: my work on this review is funded by the National Institute for Health Research (NIHR) Cochrane programme grant 13/89/16 ('Back to normal': speed and quality of recovery after surgery, major injury and critical care).

Sarah Rhodes: my salary is funded by the NIHR through three different grants, partly to provide statistical support to the CLAHRC Wounds Group.

Andrew F Smith: NIHR Cochrane Collaboration programme grant for programme of reviews in perioperative care.

SOURCES OF SUPPORT

Internal sources

- No sources of support supplied

External sources

- This project was supported by the National Institute for Health Research, via Cochrane Infrastructure funding to Cochrane Wounds. The views and opinions expressed are those of the authors and not necessarily those of the NIHR, NHS or the Department of Health and Social Care, UK.

- NIHR Cochrane Programme Grant 13/89/16 'Back to normal': speed and quality of recovery after surgery, major injury and critical care, UK.
- National Institute for Health Research Collaboration for Leadership in Applied Research and Care (NIHR CLAHRC) Greater Manchester, UK.

Sarah Rhodes was partly funded by the NIHR CLAHRC Greater Manchester. The funder had no role in the design of the studies, data collection and analysis, decision to publish, or preparation of the manuscript. However, the review may be considered to be affiliated to the work of the NIHR CLAHRC Greater Manchester. The views expressed herein are those of the authors and not necessarily those of the NHS, NIHR or the Department of Health.

DIFFERENCES BETWEEN PROTOCOL AND REVIEW

We made the following changes from the protocol ([Lewis 2016](#)).

1. New author: we added an additional author to the review (Sarah Rhodes).
2. Criteria for considering studies in this review: we did not exclude studies in which participants were diagnosed with a hospital-acquired infection (HAI) prior to randomisation. This was not reported in four studies ([Bleasdale 2007](#); [Climo 2013](#); [Milstone 2013](#); [Noto 2015](#)), such that we could not be certain whether participants in these studies had been monitored for an HAI at enrolment; and two included studies reported a small number of participants with some infections at baseline ([Camus 2005](#); [Swan 2016](#)). We included all studies, but collected baseline data on HAI as reported by study authors. However, we believe this indicated indirectness, and we downgraded the certainty of the evidence for this reason. We excluded studies of neonates because these participants have a different set of critical care needs. We specified the exclusion of studies in which only one body part was bathed; our intention was to look at the effect of chlorhexidine when used for bathing of all body areas.
3. The original protocol stated 'For studies with a cross-over design, we will only include data from the first intervention period, i.e. before cross-over to the alternative treatment'. We did not anticipate that we would identify any cluster-randomised crossover trials; however it was felt that the cluster-randomised cross-over trial was a valid design to answer the research question, given that the same participants were unlikely to be included in both the intervention and control period. We decided post-hoc to include, where possible, both periods of cluster-randomised cross-over trials, using methods described in [Unit of analysis issues](#).
4. Data synthesis: in the protocol we stated that the primary outcome (hospital-acquired infections) would be analysed using odds ratios. Most of the trials reported infections using number of events and rates rather than number of people having at least one infection; this included several cluster-randomised cross-over trials that analysed infections using rate differences with appropriately adjusted confidence intervals. In order to utilise data from as many trials as possible, and to incorporate the cluster-randomised cross-over trials, we chose to use rate differences as the summary statistic for the primary outcome.
5. Sensitivity analysis: because of the inclusion of cluster-randomised cross-over studies, and subsequent changes to analysis of data, we used sensitivity analysis to assess the inclusion of such studies. We expanded the sensitivity analysis to include analysis of a secondary outcome (mortality) because we also included cluster-randomised cross-over studies in this analysis.

INDEX TERMS

Medical Subject Headings (MeSH)

*Critical Illness; Anti-Infective Agents, Local [*therapeutic use]; Baths; Central Venous Catheters [adverse effects]; Chlorhexidine [*therapeutic use]; Cross Infection [*prevention & control]; Pneumonia, Ventilator-Associated [prevention & control]; Randomized Controlled Trials as Topic; Sepsis [prevention & control]

MeSH check words

Humans

6.4 PAPER 4

Cotterill S, Powell R, Rhodes S, Brown B, Roberts J, Tang MY, Wilkinson J. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. 2019;8:176. <https://doi.org/10.1186/s13643-019-1077-6>

PROTOCOL

Open Access



The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis

Sarah Cotterill^{1*}, Rachael Powell², Sarah Rhodes¹, Benjamin Brown^{3,4}, Jane Roberts⁵, Mei Yee Tang¹ and Jack Wilkinson¹

Abstract

Background: Health workers routinely carry out clinical behaviours, such as prescribing, test-ordering or hand-washing, which impact on patient diagnoses, care, treatment and recovery. Social norms are the implicit or explicit rules that a group uses to determine values, beliefs, attitudes and behaviours. A social norms intervention seeks to change the clinical behaviour of a target health worker by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person. This study aims to find out whether or not social norms interventions are effective ways of encouraging health workers to carry out desired behaviours and to identify which types of social norms intervention, if any, are most effective.

Methods: A systematic review will be conducted. The inclusion criteria are a population of health professionals, a social norms intervention that seeks to change a clinical behaviour, and randomised controlled trials. Searches will be undertaken in MEDLINE, EMBASE, CINAHL, British Nursing Index, ISI Web of Science, PsycINFO and Cochrane trials. Titles and abstracts will be reviewed against the inclusion criteria to exclude any that are clearly ineligible. Two reviewers will independently screen all the remaining full texts to identify relevant papers. For studies which meet our inclusion criteria, two reviewers will extract data independently, code for behaviour change techniques and assess quality using the Cochrane risk of bias tool. The primary outcome measure will be compliance with desired behaviour. To assess the effect of social norms on the behaviour of health workers, we will perform fixed effects meta-analysis and present forest plots, stratified by behaviour change technique. We will explore sources of variation using meta-regression and may use multi-component-based network meta-analysis to explore which forms of social norms are more likely to be effective, if our data meet the necessary requirements.

Discussion: The study will provide evidence regarding the effectiveness of different methods of applying social norms to change the clinical behaviour of health professionals. We will disseminate the research to academics, health workers and members of the public and use the findings from the review to plan future research on the use of social norms with health workers.

Systematic review registration: PROSPERO CRD42016045718. Future protocol changes will be clearly stated in PROSPERO.

Keywords: Systematic review, Meta-analysis, Social norm, Social comparison, Information about others' approval, Credible source, Social reward, Social incentive, Feedback, Behaviour change

* Correspondence: sarah.cotterill@manchester.ac.uk

¹Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Full list of author information is available at the end of the article



Background

Health workers routinely carry out behaviours which impact on patient diagnoses, care, treatment and recovery. Many of these behaviours have clear guidelines for best practice. Examples include appropriate ordering of diagnostic tests [1, 2], appropriate prescription of antibiotics [3, 4], regular recall of patients with long-term conditions [5], hand-washing [6] and choice of wound dressings [7]. Health workers face many challenges in following evidence-based professional practice. There is evidence that social influences are important in clinical practice [8, 9].

One proposed solution has been to implement behaviour change interventions based on social or peer norms. Social norms are the implicit or explicit rules that a group uses to determine values, beliefs, attitudes and behaviours. A social norms intervention seeks to change the clinical behaviour of a target health worker by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person. These social norms interventions can form part of an audit and feedback initiative [10–12] or may be developed as another behaviour change intervention [13]. These are interventions with reach that can be implemented routinely across multiple health workers and settings at low cost, so the absolute gain can be very large. We use the term target to refer to a health worker who is targeted by social norms interventions, with a view to changing their clinical behaviour. We use the term reference group or reference person to mean a person or group of people used as a reference category in a social norms intervention. For the purposes of our review, we anticipate that reference categories will include people with the same profession or occupation as the target; people employed by the same organisation as the target; people who deliver, administer, manage, commission or make policy on health services; or professional bodies such as royal colleges and trade unions. It is possible that some studies will use social norms approaches where the reference group is not taken from the above list (such as credible source from a celebrity or exposing the target's behaviour to patients). We will include in the review papers with any type of reference group.

The ability of social norms to affect behaviour has been considered within several behaviour change theories and theoretical frameworks. For example, 'subjective norm' is a construct within the Theory of Planned Behaviour [14], which describes an individual's perceptions of whether valued others think one should perform a behaviour, combined with one's motivation to comply with others' beliefs. The Theory of Normative Social Behaviour [15] proposes that behaviour can be changed through normative mechanisms and has made distinctions between descriptive norms (beliefs concerning the prevalence of a behaviour) and injunctive norms (beliefs concerning what one feels they ought to do based on others' expectations—social approval).

Further, the 'social influences' domain of the Theoretical Domains Framework [16] also includes several normative constructs: social norms, social comparisons and group norms. We will include studies based on either descriptive norms or injunctive norms messages. A descriptive norms message provides the target with information about the behaviour of others in the reference group. Examples of descriptive norms interventions include giving the target information about the behaviour of a reference person or group or comparing the target's behaviour with the behaviours of a reference person or group. An injunctive norms message provides the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval. Examples of injunctive norms interventions include providing the target with information about whether the behaviour has the approval/disapproval of the reference group or person, exposure (actual or promised) of the target's behaviour to a reference group and praise, commendation, applause or thanks (actual or promised) from a reference group or person.

The behaviour change technique taxonomy v1 is a list of 93 distinct behaviour change techniques (BCTs) which are used in behaviour change interventions [17]. The BCT Taxonomy includes five BCTs which we believe involve social norms: social comparison, information about others' approval, credible source, social reward and social incentive [17]. We have chosen to define social norms in terms of the BCT taxonomy v1 because, based on international consensus, it aims to define and label all active ingredients of interventions, including social norms. It incorporates previous behaviour change taxonomies and has involved significant effort from leaders in the field and considerable investment from the MRC and NIHR in developing the taxonomy. We believe this to be the most reliable tool currently available that can define BCTs. We have selected the five BCTs that we consider have a social norms element to them, and we have discussed this selection carefully, both within the research team and with our steering group of international experts. We are open to the possibility that studies may be eligible for the review that test social norms interventions but do not incorporate one of these five identified BCTs.

Health workers frequently receive audit and feedback (A&F), which involves 'providing a recipient with a summary of their performance over a specified period of time' ([10] p. 1). Social norms interventions are sometimes included as one component of A&F, such as when the health worker is shown information about their own performance and also a comparison with their peers [11, 12]. A&F has already been shown to be effective in changing health worker behaviour, but with large variation in outcomes depending on the context and the intervention design [18]. There is a need to understand the ingredients for successful A&F [10, 19], and the effects or mechanisms of the social norms constituents of A&F have been identified in a recent

systematic review as topics for further research [10]. Our review will contribute to this important research agenda by systematically examining the evidence for using social norms BCTs with health workers.

Aims

The overall aim is to conduct a systematic review to assess, among health workers, the impact of social norms BCTs, compared to alternative interventions, no intervention or comparison of one or more social norms BCTs on compliance with evidence-based professional practice. The review will address two research questions:

1. What is the effect of social norms interventions on the clinical behaviour of health workers and resulting patient outcomes?
2. Which contexts, modes of delivery and behaviour change techniques are associated with the effectiveness of social norms interventions on health worker clinical behaviour change?

Methods

This protocol follows the PRISMA-P reporting guidelines for systematic reviews [20] (PRISMA-P checklist included as Additional file 1).

Eligibility criteria

The inclusion criteria for the review are a population of health professionals, a social norms intervention that seeks to change a clinical behaviour, and the study type is a randomised controlled trial.

Population

The population of interest is health workers and managers. Student health workers will be included, but only if the study is in a healthcare setting. Any healthcare setting will be eligible, including care homes, nursing homes and patients' own homes. Interventions in educational establishments or simulated environments will not be eligible.

Interventions

The systematic review will focus on social norms interventions, defined as interventions seeking to change the clinical behaviour of a target health worker by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person. We have selected five BCTs (from the BCT taxonomy v1) that we consider have a social norms element to them (6.2. Social comparison; 6.3. Information about others' approval; 9.1 Credible source; 10.4 Social reward; 10.5 Social incentive), but we are open to the possibility that studies may be eligible for the review that test social norms interventions without using one of the five BCTs. Three BCTs are used unchanged in this review (social comparison, information about others'

approval and credible source). Two BCTs have been adapted slightly for clarity: the definitions of social reward as 'verbal or non-verbal reward' and social incentive as 'verbal or non-verbal incentive' are insufficient to distinguish a 'social' reward incentive from other types of reward or incentive. Further, in the present study, we are interested in only those social rewards or incentives that rely on social norms. We define social reward and incentive as involving praise, commendation, applause or thanks, all of which are injunctive norms messages, providing the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval (Table 1).

Included studies must state a behaviour that is being targeted for change. By definition, BCTs relate to behaviour(s): 'a single action or sequence of actions'. Either the 'performance of wanted behaviour(s) and/or inhibition (non-performance) of unwanted behaviour(s)' might be addressed by a BCT [17] (detail/quotes are from electronic supplementary materials, p1.). We will report the number of studies which would otherwise meet our inclusion criteria but do not mention a target behaviour.

The format of the behaviour change intervention may be letter, electronic or verbal. It may be delivered once only, repeated over time or delivered in a timely fashion on occasions when the behaviour is expected to be performed. For example, an intervention to reduce prescribing of antibiotics by family doctors might be delivered once only, by regular weekly email, or by a computerised reminder when a relevant disease code is entered into the practice computer system.

Comparators

We anticipate finding a range of comparators, including alternative intervention, no intervention or comparison of one or more social norms BCTs (Table 2).

Study designs

The systematic review will only include randomised controlled trials (RCTs) of any design (cluster, factorial, parallel, cross-over and stepped wedge). The justification for restricting the review to RCTs is that the review is concerned with the effectiveness of social norms, and randomised controlled trials are the best method for assessing the effectiveness of an intervention. We will include both published and unpublished research. Studies must be reported in English because the research team has no resource for translation from other languages.

Information Sources and search strategy

The search strategy was developed collaboratively between the researchers and the information specialist in our review team. The search targeted databases relevant to health, social and behavioural science, without restriction on dates: MEDLINE, Ovid; EMBASE, Ovid; CINAHL, Ebsco; British

Table 1 Social norms BCTs for inclusion in the review

Name and Definition from BCT Taxonomy [17]	SOCIAL review name and definition
<p>6.2. Social Comparison Draw attention to others' performance to allow comparison with the person's own performance. Note: being in a group setting does not necessarily mean that social comparison is actually taking place. Example: Show the doctor the proportion of patients who were prescribed antibiotics for a common cold by other doctors and compare with their own data.</p>	<p><i>6.2. Social Comparison—unchanged</i></p>
<p>6.3. Information about others' approval Provide information about what other people think about the behaviour. The information clarifies whether others will like, approve or disapprove of what the person is doing or will do. Example: Tell the staff at the hospital ward that staff at all other wards approve of washing their hands according to the guidelines.</p>	<p><i>6.3. Information about others' approval—unchanged</i></p>
<p>9.1. Credible source Present verbal or visual communication from a credible source in favour of or against the behaviour. Note: code this BCT if source generally agreed on as credible, e.g. health professionals, celebrities or words used to indicate expertise or leader in field and if the communication has the aim of persuading. Example: Present a speech given by a high status professional to emphasise the importance of not exposing patients to unnecessary radiation by ordering X-rays for back pain.</p>	<p><i>9.1. Credible source—unchanged</i></p>
<p>10.4. Social reward Arrange verbal or non-verbal reward if and only if there has been effort and/or progress in performing the behaviour (includes 'Positive reinforcement'). Example: Congratulate the person for each day they eat a reduced fat diet.</p>	<p><i>10.4. Social reward—changed</i> Arrange praise, commendation, applause or thanks if and only if there has been effort and/or progress in performing the behaviour (includes 'Positive reinforcement'). Example: Arrange for a family doctor to be sent a thank you note for each week that they reduce their level of antibiotic prescribing. Reason for change: the definition of social reward as 'verbal or non-verbal reward' is insufficient to distinguish a 'social' reward from other types of reward. Further, in the present study, we are interested in only those social rewards that rely on social norms. Praise, commendation, applause or thanks are all injunctive norms messages, providing the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval.</p>
<p>10.5 Social incentive Inform that a verbal or non-verbal reward will be delivered if and only if there has been effort and/or progress in performing the behaviour (includes 'Positive reinforcement'). Example: Inform that they will be congratulated for each day that they eat a reduced fat diet.</p>	<p><i>10.5 Social incentive—changed</i> Inform that praise, commendation, applause or thanks will be delivered if and only if there has been effort and/or progress in performing the behaviour (includes 'Positive reinforcement'). Example: Promise a family doctor in advance that they will be sent a thank you note for each week that they reduce their level of antibiotic prescribing. Reason for change: the definition of social reward as 'verbal or non-verbal reward' is insufficient to distinguish a 'social' reward from other types of reward. Further, in the present study, we are interested in only those social rewards that rely on social norms. Praise, commendation, applause or thanks are all injunctive norms messages, providing the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval.</p>

Table 2 Types of comparison

	Interventions		Controls
1	Social norm intervention	vs	Any control
2	Social norm intervention + X	vs	X
3	Social norm intervention + X	vs	Any control
4	Social norm intervention + X	vs	Social norm intervention
5	Social norm intervention A	vs	Social norm intervention B

Where X is any other intervention and A and B are two different types of social norm behaviour change technique.

Nursing Index; ISI Web of Science; PsycINFO and Cochrane trials. The search was structured to find the populations (health workers), interventions (social norms) and study types (RCTs) of interest. The population search terms were based on health worker search terms from previous reviews, supplemented by a list of health worker roles from a local NHS trust and review by the study team. Intervention search terms were based on the descriptions of social norms BCTs in the BCT taxonomy v1, audit and feedback search terms, and theories relevant to social norms. RCT search filters are those described in Chapter 6.4,11 of the Cochrane Handbook for Systematic Reviews of

Interventions [21]. The search terms were developed by the study team and used to develop searches in MEDLINE. They were then reviewed by the project team and translated into the other databases using the appropriate controlled vocabulary as applicable. The final search strategy was reviewed by the Management Group and Steering Group. Searches were completed between 4 and 19 July 2018. The MEDLINE search is available within our PROSPERO registration [22] and as Additional file 2.

Data collection

Data management

Covidence will be utilised as a management tool for the review (<https://www.covidence.org/>).

Study selection process

One reviewer will independently screen the titles and abstracts and exclude studies which obviously do not meet inclusion criteria. A second reviewer will independently screen a sample of 20% of records. If there is any difference of opinion at the title/abstract stage, the reviewers will err on the side of inclusion. The number of cases of disagreement will be reported. If the level of disagreement is over 10%, all the records will be double screened. Two reviewers will independently screen the full texts and apply the eligibility criteria. If there is any difference of opinion, the two screeners will discuss, and if they cannot agree, the text will be reviewed by a third member of the research team or resolved at a project team meeting if required.

Data collection process

Data from included studies will be extracted independently by two researchers. A data collection template, based on the Cochrane EPOC data collection form [23] has been developed [22], which will ensure that (where available) information on population and setting, methods, participants, interventions, controls, outcomes, results and applicability is collected. Discrepancies between reviewers will be resolved through discussion between the two researchers, by involving a third member of the research team, or discussion within the full project team meeting where necessary. Interventions will be described using relevant items from the TIDieR checklist [24] and specific behaviour change techniques will be classified using the behaviour change technique taxonomy v1 [17].

Unit of analysis issues

If any of the studies in the review are cluster randomised trials, summary measures (e.g. means, odds ratios) and adjusted standard errors will be extracted from appropriately analysed trials. Where necessary, adjustments for clustering will be made, using the ICC (intra-class correlation coefficient). If more than one comparison from a study with more than two arms is eligible for the same comparison,

the number of health workers in the shared arm will be adjusted to avoid double counting. The adjustment will be done by dividing the number of health workers in the shared arm approximately evenly among the comparisons.

Missing data

The research team will search for companion papers (by author searching and citation searching) and/or contact trial authors once to obtain missing information under the following circumstances:

1. Where there is insufficient information in the full trial report to establish whether or not the trial meets our inclusion criteria
2. Where it is clear that our primary outcome measure was measured but insufficient information was reported to establish the number of participants and/or summary measures
3. Where the intervention descriptions are insufficiently clear to determine whether or not the trial meets our inclusion criteria

The research team will impute estimates of standard deviations where necessary (after contacting authors) using standard deviations from other similar studies (same target behaviour, same setting) that use the same type of outcome. Where necessary, for cluster randomised trials, a value of the ICC from similar studies (same target behaviour, same outcome measure, same setting) will be imputed.

Piloting

All processes for screening, extraction, BCT coding and assessment of bias will be piloted within the team prior to implementation.

Coding of Behaviour Change Techniques

Specific behaviour change techniques will be coded independently by two researchers, using the behaviour change technique taxonomy v1 [17]. The reliability of identifying and coding behaviour change techniques from intervention descriptions has been assessed [25]. This research assessed the reliability of judgements between different coders and across time using the prevalence and bias-adjusted kappa (PABAK) statistic. Overall, there were high rates of reliability between coders. Of the five BCTs relevant to this review, three of the techniques (social comparison, information about others' approval, social incentive) were assessed in a high number of studies (6 to 20 studies) and were found to have high inter-rater reliability (over 0.7). Social reward was assessed in only one study but the reliability was high (1.0). Credible source was found in a high number of studies but found to have lower reliability (0.4) (Table 3).

All coders will be trained in coding of behaviour change techniques, using the on-line training provided by the

authors of the BCT taxonomy (<http://www.bct-taxonomy.com/>), which is required to meet acceptable standards of competence. The online training has been shown to improve agreement with expert consensus, confidence for BCTs assessed and coding competence [26]. Additionally, they have attended a workshop facilitated by co-applicant RP and steering group member MJ. The level of agreement on the coding of BCTs between two coders will be reported, using a PABAK statistic.

Outcomes and prioritisation

The primary outcome for this review is compliance of the health worker with the desired clinical behaviour (e.g. rate of antibiotic use) at 6 months post randomisation. Six months post randomisation was chosen to identify a common time point from randomisation, and there is the potential for interventions to run over several months. Researchers will extract details of the outcome closest to 6 months post randomisation (e.g. mean and standard deviation or proportion). Other time points will be noted for inclusion in the description of studies and the team may consider conducting analysis using earlier/later time points, where reported.

It is likely that different studies will use different outcome measures as they will be measuring different behaviours and using different methods to assess those behaviours. The research team will convert any observed measure of health worker behaviour into a standardised mean difference between groups in terms of compliance with the desired behaviour. Examples include the mean number of times the behaviour was performed per worker or the mean rate of behaviour (e.g. rate of antibiotic items dispensed per 1000 population). It is possible that compliance will be reported as a binary outcome, for example compliance vs non-compliance on a single occasion, e.g. attendance a training session. The methods of Chin 2000 will be adopted to convert binary outcomes to standardised mean differences with associated standard errors using the formula [27]:

$$\text{SMD} = \frac{\sqrt{3}}{\pi} \ln \text{OR}$$

If several measures of compliance are reported in a trial, the following criteria will be used to select the outcome for

Table 3 Inter-rater reliability of coding of social norms BCTs

	Behaviour change technique	Number of studies in which the BCT was present	Inter-rater reliability (PABAK)
6.2	Social comparison	13	0.76
6.3	Information about others' approval	6	0.94
9.1	Credible source	32	0.4
10.4	Social incentive	7	0.9
10.5	Social reward	1	1.0

Source: [24]

the primary analysis, in decreasing order of importance: (a) observed measure rather than self-report, (b) continuous measure, (c) final score rather than change from baseline or percentage change, (d) described as the primary outcome, (e) used to calculate the sample size, and (f) reported first.

A secondary outcome for the review is any patient outcomes which are likely to result from targeting the health worker behaviour.

Risk of bias in individual studies

Two reviewers will independently assess the risk of bias of each study; discrepancies will be resolved by discussion between the two reviewers, by involving a third reviewer, or by discussion within the project team, if needed. The risk of bias for each main outcome in all studies included in the review will be assessed using the tool described in Chapter 8 of the Cochrane Handbook for Systematic Reviews of Interventions [28]. An assessment of the risk of bias (high, low or unclear risk of bias) on each domain (random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting and other bias) will be assigned to each of the included studies and will be reported and utilised in sensitivity analysis.

Data synthesis

Criteria for study data to be meta-analysed

The meta-analysis will only include those studies that report a relevant outcome measure (clinical behaviour of a health worker or patient health outcome) that can be converted into a standardised mean difference.

Planned approach for meta-analysis

To address the first research question (What is the effect of social norms interventions on the clinical behaviour of health workers, and resulting patient outcomes?) estimates from the individual studies will be combined using a fixed effects meta-analysis on the standardised mean difference, stratified by the 5 social norms BCTs. The research team prefers a fixed effects approach to a random effects approach in this case, since a key assumption of the latter, exchangeability, is not anticipated to hold in these trials [29]. The fixed effects analysis will yield a summary of the evidence in these trials, rather than an estimate of a common underlying treatment effect, as advocated by Higgins et al. [30]. Statistical heterogeneity will be explored and reported visually by preparing forest plots and reporting I^2 .

To address the second research question (Which contexts, modes of delivery and behaviour change techniques are associated with the effectiveness of social norms interventions on health worker clinical behaviour

change?), if there are sufficient studies with comparable outcomes, the research team will follow steps 1 to 3:

Step 1: Explore sources of variation, using forest plots and narrative description.

Step 2: Undertake an exploratory analysis, using multi-variable meta-regression to investigate sources of heterogeneity and explain variation in the results. Meta-regression is an appropriate method in which appropriate weights are assigned to studies/sub-groups. Prior to undertaking this analysis, a detailed analysis plan will be written, making explicit the sources of variation that will be investigated and our hypotheses, following the recommendations of Thompson [31]. Our categorisation of the sources of variation was informed by a meta-synthesis of audit and feedback interventions [32]. Sources of variation may include the items

1. Context

- Type of health worker (such as a family doctor, nurse, secondary care doctor or allied health professional).
- Type of behaviour (such as prescribing, hand-washing or surgical technique).
- Any targeting of participants based on baseline performance of behaviour (such as below or above average).
- Concomitant behaviour change techniques delivered alongside the social norms intervention
- Choice of reference group (such as health worker, professional body, patient or other)
- Direction of change expected (increase, decrease or maintenance of behaviour)

2. Mode of delivery

- Who delivers the intervention (such as health worker, non-health worker, patient or researcher) and whether they are internal or external to the target's organisation.
- Frequency and intensity
- Delivery method (such as email, letter, computerised or face-to-face)

3. Social norm behaviour change technique

Step 3: Explore whether social norms interventions with particular components and concomitant behaviour change techniques are more likely to be effective. For this, the research team will consider using a multi-component-based network meta-analysis [33]. This analysis will be reliant on 2 conditions: (1) being able to identify distinct components/techniques from the published literature (2) a connected network of components/techniques. The research team will only proceed with this analysis if these two conditions hold and a pre-specified analysis plan is approved by the Study Steering Committee.

Additional analyses

Sensitivity analyses for the primary outcome (a) include only studies with a low risk of bias (for each separate domain and all domains), (b) include only studies where the primary outcome was reported on a continuous scale, (c) using methods of Ma et al. [34] to impute missing standard deviations, and (d) include only studies where the standard error was not imputed.

Planned summary if meta-analysis is not possible

If meta-analysis is not possible because of inconsistency and incomplete reporting of outcome measures, an Albatross plot, as proposed by Harrison, will be produced [35]. In an albatross plot, p values, ordered from extreme negative trend to extreme positive trend, are plotted against study size. Effect contours will be added to show a range of effect sizes.

Meta-bias

The impact of reporting bias will be minimised by performing a comprehensive search for eligible studies. Publication bias in the reported studies will be investigated using a funnel plot.

Confidence in cumulative evidence

The strength of the body of evidence will be assessed for the primary outcome using GRADE for each of the five BCTs and for social norms overall.

Discussion

There are many implementation research contexts in which modification of the behaviour of health workers may have a beneficial effect on patient diagnosis, care, treatment, and on the costs of healthcare. These contexts include situations where health workers are expected to follow evidence-based professional practice such as prescribing, ordering tests, choosing treatments and adhering to guidelines. A systematic review of the evidence is needed to establish whether these interventions are effective and what factors influence their effectiveness. Limitations of the review include the following: the search strategy may not pick up every healthcare profession and could miss social norms interventions that are described using bespoke terminology, the context in which clinical behaviour takes place and the factors that influence it are complex and it is unlikely the individual studies will report these fully and consistently, the review will synthesize the results of trials with different outcome measures, by restricting the studies to those written in English, we may miss important evidence, and authors will not be contacted for intervention manuals which could lead to the omission or misclassification of some BCTs. All this could have a potential impact on interpretation of the results and

recommendations for future research. This is the first systematic review, to our knowledge, that will investigate the effect of social norms interventions on health worker behaviour and resulting patient outcomes.

Additional files

Additional file 1: PRISMA-P Checklist. (DOCX 34 kb)

Additional file 2: Final search strategies. (DOCX 31 kb)

Abbreviations

A&F: Audit and feedback; BCT: Behaviour change technique; EPOC: Effective Practice and Organisation of Care; GRADE: Grading of Recommendations, Assessment, Development and Evaluations; ICC: Intra-class correlation coefficient; MRC: Medical Research Council; NHS: National Health Service; RCT: Randomised controlled trial; TIDieR: Template for Intervention Description and Replication

Acknowledgements

The authors thank the experts on our study steering group who have provided valuable advice and guidance: Sofia Dias, Robbie Foy, Marie Johnston and Manoj Mistry.

Authors' contributions

SC is the guarantor. SC drafted the protocol. SC, RP, SR and BB wrote the funding bid on which this protocol is based. SC, RP, SR, BB, MYT and JW contributed to the development of the selection criteria, the risk of bias assessment strategy and data extraction criteria. SC, RP, MYT and JR developed the search strategy. SR and JW provided statistical expertise. All authors read, provided feedback and approved the final manuscript.

Funding

This report is independent research funded by the National Institute for Health Research (Health Services and Delivery Research, 17/06/06, The impact of social norms interventions on clinical behaviour change among health workers: a systematic review). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the sponsor (The University of Manchester).

Availability of data and materials

Not applicable

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester M13 9PL, UK. ²Manchester Centre for Health Psychology, Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester, UK. ³Health e-Research Centre, Farr Institute for Health Informatics Research, Faculty of Biology Medicine and Health, University of Manchester, Manchester, UK. ⁴Centre for Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester, UK. ⁵Outreach and Evidence Search Service, Library & E-learning Service, The Pennine Acute Hospitals NHS Trust, Royal Oldham Hospital, Oldham, UK.

Received: 27 September 2018 Accepted: 24 June 2019

Published online: 18 July 2019

References

- Hamilton W, Watson J, Round A. Investigating fatigue in primary care. *Br Med J*. 2010;341.
- NICE. Cardiovascular disease: risk assessment and reduction, including lipid modification Clinical guideline [CG181] Published July 2014, updated September 2016 [nice.org.uk/guidance/cg181](https://www.nice.org.uk/guidance/cg181). 2014.
- Davies SC, Fowler T, Watson J, Livermore DM, Walker D. Annual Report of the Chief Medical Officer: infection and the rise of antimicrobial resistance. *Lancet*. 2013;381:1606–9.
- Hawker JI, Smith S, Smith GE, Morbey R, Johnson AP, Fleming DM, Shallcross L, Hayward AC. Trends in antibiotic prescribing in primary care for clinical syndromes subject to national recommendations to reduce antibiotic resistance, UK 1995–2011: analysis of a large database of primary care consultations. *J Antimicrob Chemother*. 2014;69:3423–30.
- NICE. Guidance. Type 2 diabetes in adults: management. Published: 2 December 2015. [nice.org.uk/guidance/ng28](https://www.nice.org.uk/guidance/ng28).
- Loveday HP, Wilson JA, Pratt RJ, Golsorkhi M, Tingle A, Bak A, Browne J, Prieto J, Wilcox M. epic3: National Evidence-Based Guidelines for Preventing Healthcare-Associated Infections in NHS Hospitals in England. *Journal of Hospital Infection*. 86:S1–S70.
- NICE advice. Chronic wounds: advanced wound dressings and antimicrobial dressings. In: Evidence summary, Published: 30 March 2016. [nice.org.uk/guidance/esmpb2](https://www.nice.org.uk/guidance/esmpb2).
- Whiting P, Toerien M, de Salis I, Sterne JAC, Dieppe P, Egger M, Fahey T. A review identifies and classifies reasons for ordering diagnostic tests. *Journal of Clinical Epidemiology*. 2007;60:981–9.
- van der Weijden T, van Bokhoven MA, Dinant G-J, van Hasselt CM, Grol RPTM. Understanding laboratory testing in diagnostic uncertainty: a qualitative study in general practice. *Br J Gen Pract*. 2002;52:974–80.
- Ivers NM, Sales A, Colquhoun H, Michie S, Foy R, Francis JJ, Grimshaw JM. No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention. *Implement Sci*. 2014;9:14.
- Ivers NM, Tu K, Young J, Francis JJ, Barnsley J, Shah BR, Upshur RE, Moineddin R, Grimshaw JM, Zwarenstein M. Feedback GAP: pragmatic, cluster-randomized trial of goal setting and action plans to increase the effectiveness of audit and feedback interventions in primary care. *Implementation Science*. 2013;8:142.
- Carney PA, Abraham L, Cook A, Feig SA, Sickles EA, Miglioretti DL, Geller BM, Yankaskas BC, Elmore JG. Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. *Academic Radiology*. 2012;19:1114–20.
- Huis A, Schoonhoven L, Grol R, Donders R, Hulscher M, van Achterberg T. Impact of a team and leaders-directed strategy to improve nurses' adherence to hand hygiene guidelines: a cluster randomised trial. *Int J Nurs Stud*. 2013;50:464–74.
- Ajzen I. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*. 1991;50:179–211.
- Rimal RN, Real K. How Behaviors are Influenced by Perceived Norms: A Test of the Theory of Normative Social Behavior. *Commun Res*. 2005;32:389–414.
- Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science*. 2012;7:37.
- Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Eccles MP, Cane J, Wood CE. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med*. 2013;46:81–95.
- Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, O'Brien MA, Johansen M, Grimshaw J, Oxman AD. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev*. 2012;(6):CD000259.
- Gardner B, Whittington C, McAteer J, Eccles MP, Michie S. Using theory to synthesise evidence from behaviour change interventions: the example of audit and feedback. *Soc Sci Med*. 2010;70:1618–25.
- Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015;4(1).

21. Higgins JPT GSe. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]; Cochrane Collab; 2011.
22. Cotterill S, Powell R, Rhodes S, Roberts J, Tang MY, Wilkinson J: The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review. PROSPERO 2016 CRD42016042718 Available from: https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=42718.
23. Effective Practice and Organisation of Care (EPOC). Data collection form: EPOC resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services; 2013.
24. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, Altman DG, Barbour V, Macdonald H, Johnston M, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014;348:g1687.
25. Abraham C, Wood CE, Johnston M, Francis J, Hardeman W, Richardson M, Michie S. Reliability of Identification of Behavior Change Techniques in Intervention Descriptions. *Ann Behav Med*. 2015;49:885–900.
26. Wood CE, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, Michie S. Applying the behaviour change technique (BCT) taxonomy v1: a study of coder training. *Translational Behavioral Medicine*. 2015;5:134–48.
27. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*. 2000;19:3127–31.
28. Higgins JPT AD, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* (updated March 2011). Edited by Higgins JPT GSe: The Cochrane Collaboration; 2011.
29. Charles Poole, Sander Greenland, Random-Effects Meta-Analyses Are Not Always Conservative, *American Journal of Epidemiology*. 1999;150(5): 469–75. <https://doi.org/10.1093/oxfordjournals.aje.a010035>
30. Higgins JPT, López-López JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM, McGuinness LA, Moore THM, Rehfues EA, Thomas J, Caldwell DM. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health*. 2019;4:e000858.
31. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21:1559–73.
32. Benjamin Brown, Wouter T. Gude, Thomas Blakeman, Sabine N. van der Veer, Noah Ivers, Jill J. Francis, Fabiana Lorencatto, Justin Pesseau, Niels Peek and Gavin Daker-White. Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implementation Science* 2019 14:40 <https://doi.org/10.1186/s13012-019-0883-5>
33. Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. *Evidence Based Mental Health*. 2016.
34. Ma J, Liu W, Hunter A, Zhang W. Performing meta-analysis with incomplete statistical information in clinical trials. *BMC Medical Research Methodology*. 2008;8:56.
35. Harrison S. Albatross plots, a novel method of displaying data for otherwise un-combinable studies. In: *Young Statisticians Meeting*. UK: Cardiff University; 2015.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



6.5 PAPER 5

Cotterill S, Tang MY, Powell R, Howarth E, McGowan L, Roberts J, Brown B, Rhodes S. Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis. *Health Serv Deliv Res* 2020;8:41, pp 35-77.

<https://doi.org/10.3310/hsdr08410>

(Methods and Results chapters included only).

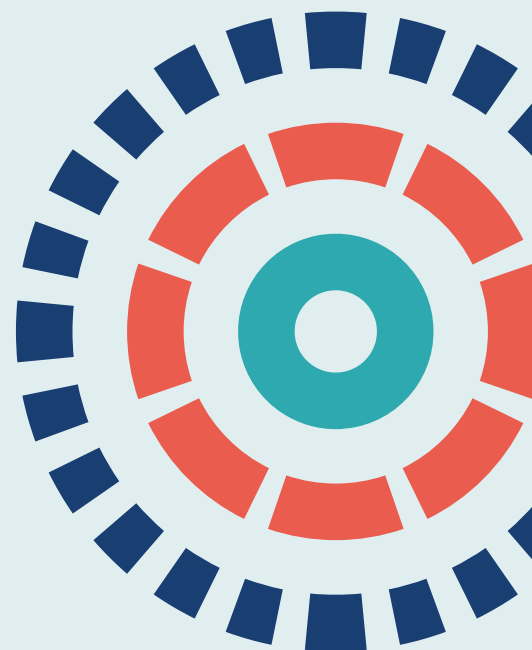
Health Services and Delivery Research

Volume 8 • Issue 41 • October 2020

ISSN 2050-4349

Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis

Sarah Cotterill, Mei Yee Tang, Rachael Powell, Elizabeth Howarth, Laura McGowan, Jane Roberts, Benjamin Brown and Sarah Rhodes



Chapter 2 Methods

Parts of this chapter have been reproduced from our review protocol.^{1,2} This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>. The text below includes minor additions and formatting changes to the original text.

Changes to the protocol

The published protocol differs from the funding proposal in the ways listed below, and these changes were approved by NIHR during the early months of the project:

- Change in terminology – ‘social norms’ replaces ‘social influence’. The justification for this was twofold. First, the term ‘social influence’ is a domain within the Theoretical Domains Framework²³ and encompasses a broad range of social concepts, such as emotional and practical support, demonstrating a behaviour and changing the social environment, as well as social norms; it is not specific enough for the purpose of this review. Second, ‘social norms’ better captures the core mechanism by which we expected the interventions to have an effect.
- We added to our inclusion criteria a requirement that included studies must state a behaviour that is being targeted for change. This was not a fundamental change from the earlier version, but was stated more clearly than previously.
- Change from ‘health professional’ to ‘health worker’. This is a clarification rather than a change to the original inclusion criteria. It was always our intention to include all staff providing health care, and this change of terminology makes clear that not all health workers have professional qualifications.

Inclusion criteria

The inclusion criteria for the review were based on the population, types of intervention and study designs, as follows.

Population

The population of interest was health workers (including managers) responsible for patient care in a health-care setting. Health workers in training were included, but only if they were in a health-care setting (i.e. not in campus or laboratory environments). Any health-care setting was eligible, including primary care, secondary care, care homes, nursing homes and patients’ own homes. Interventions taking place in simulated environments were not eligible for inclusion.

Interventions

A social norms intervention seeks to change the clinical behaviour of a target health worker by exposing them to the values, beliefs, attitudes or behaviours of a reference person or group. We looked for the five BCTs that we considered to have a social norms element to them: 6.2. social comparison, 6.3. information about others’ approval, 9.1. credible source, 10.4. social reward and 10.5. social incentive. However, we were open to including studies that met all other inclusion criteria and had a social norms element, even if they did not include any of these five BCTs.

Included studies must have stated a clinical behaviour of health workers that was targeted for change through the use of social norms. If the behaviour was not specified, it was not possible to determine which aspects of an intervention were relevant to the anticipated behaviour change. Indeed, the BCT taxonomy v1 coding guidance states that the target behaviour needs to be specified and BCTs must

target that behaviour for BCTs to be coded.²⁸ Clinical behaviour here is defined as any behaviour that is performed within a (non-simulated) environment that affects patient diagnosis, care, treatment or recovery. We have reported the number of studies identified by our search that met all other inclusion criteria but did not mention a target behaviour.

Comparators

All comparators were eligible for inclusion, including alternative interventions, no intervention or comparison of one social norms BCT with one or more other social norms BCTs.

Study designs

Only randomised controlled trials (RCTs) were included in the review. All designs of RCTs (cluster, factorial, parallel, cross over and stepped wedge) were eligible for inclusion. The justification for restricting the review to RCTs was that the review is concerned with the effectiveness of social norms, and RCTs are the best method for assessing the effectiveness of an intervention.

We included both published and unpublished research. Studies had to be reported in English because the research team had no resource for translation from other languages.

Search strategy

The search strategy was developed using an extensive iterative scoping process, involving the whole team including an information specialist (Jane Roberts). Lists of possible search terms were suggested by team members; these were developed into search strategies by Jane Roberts, who then ran preliminary searches. A sample of the titles and abstracts were reviewed closely by Sarah Cotterill and Mei Yee Tang and discussed by the wider team. This review involved consideration of whether searches were too inclusive or too restrictive, and examination of resulting abstracts to look for potential additional search terms.

The searches were based on three groups of terms: population, interventions and study design.

Population

A list of population terms was developed by looking at Cochrane reviews^{25,29} that included a similar population of health workers, augmented by job roles included in the UK national workforce data set produced by NHS Digital.³⁰

Interventions

Social norms interventions are not described consistently in the literature, and different terms are used in various academic disciplines. This presented us with the challenge of finding appropriate search terms to make sure that we would discover the full range of literature on this topic. We were aware that many studies involving A&F contain a social comparison element;²⁵ therefore, we looked at the search terms that were used in a previous systematic review of A&F.²⁵ We omitted anything relating solely to 'audit', as this was not relevant for this review.

During the scoping phase, various feedback terms were tried out. The use of 'feedback' alone produced many irrelevant papers, such as those relating to educational feedback and electronic feedback. The final search, following extensive trial and error in the piloting phase, included 'feedback' when used alongside other relevant terms (audit, monitoring, peer, performance, data, individualised, web, personalised, comparative, team, practitioner, practice and clinical or social). We also included 'benchmark'. We included some overall terms that are used in the literature on social norms: 'norm' used close to 'social', 'descriptive', 'peer' or 'subjective'; 'social influence'; 'benchmarking'; 'social or peer comparison'; and 'social competition'. Terms that appeared in behavioural economics literature were included: 'social proof', 'image motivation' and 'warm glow'.

Additional search terms were developed for each of the five social norms BCTs by looking at the text used to describe them in the BCT taxonomy v1,^{24,31} extensive discussion in the team and examining relevant articles. Additional terms for information about others' approval and credible source included 'positive reinforcement', 'congratulate', 'praise' and 'commendation'. Terms for social reward and social incentive included 'social', 'verbal' and 'non-verbal' alongside 'incentive' or 'reward'. Finally, the search included terms to describe theories that are used to explain interventions based on social norms: the Theory of Planned Behaviour,³² the Theory of Reasoned Action,³³ the Theoretical Domains Framework,³⁴ Social Cognitive Theory,³⁵ and the Theory of Normative Social Behaviour.²²

Study design

The search for RCTs was taken directly from the Cochrane RCT search described in the *Cochrane Handbook for Systematic Reviews of Interventions*.³⁶ This was translated into other relevant databases.

The search was developed in MEDLINE and then adapted for other databases. Terms relating to the same concept (e.g. different types of health workers) were combined using the Boolean operator 'OR' and different concepts (e.g. health workers and social norms) were combined using 'AND'. The search strategy was tailored for the different electronic databases using medical subject headings (MeSH) where appropriate, wildcard symbols and truncations (see *Appendix 1*). Backward- and forward-citation searching was not conducted owing to time and resource constraints.

Published literature was systematically searched on 24 July 2018 in electronic databases relevant to health and social care: Ovid MEDLINE, Ovid EMBASE, Healthcare Databases Advanced Search (HDAS) – Cumulative Index to Nursing and Allied Health Literature (CINAHL), HDAS British Nursing Index (BNI), Cochrane Central Register of Controlled Trials (CENTRAL), Ovid PsycINFO and Web of Science (see *Appendix 1* for search strategies and *Appendix 1, Table 15*, for the results).

Data collection

Study selection

The process for identifying studies for review followed the stages of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement.³⁷

All references generated from the search were managed in Covidence (Melbourne, VIC, Australia): an online screening and data extraction tool for systematic reviews. All reviewers were provided with instructions for both the title and the abstract, and full-text screening stages (see *Appendix 2*). At the title and abstract screening stage there was an initial learning phase (305 studies), during which the coders worked steadily through the task, applying the inclusion criteria to the papers and stopping after small batches to discuss any discrepancies as they went along. Disagreements and uncertainties were discussed with the wider research team. This process enabled the main coder to build up a high level of consistency. For the remaining studies, one reviewer (MYT) independently screened all of the titles and abstracts against the inclusion criteria, and another researcher (SC, SR or JW) screened a sample of 20% of the records. These were randomly selected using a random integer generator (www.random.org; accessed 1 September 2020). By the time that 20% of the records (493 studies) had been screened, there was very little difference between the decisions of the two coders, and we were confident that the main coder could make the exclusion decisions with reliability. Inter-rater reliability on these 493 studies was good³⁸ ($\kappa = 0.68$). Where there was any hesitation on her part about whether to include or exclude, she erred on the side of inclusion and continued to discuss any uncertainties with the wider team.

All of the studies at the full-text stage were independently screened by two researchers from the screening team (MYT, SC or SR). The two reviewers screened the papers concurrently using Covidence, and were not aware of the other person's recommendation until after they had entered

their own. The screening involved reading the full-text paper and deciding whether or not the paper met the eligibility criteria (a population of health workers in a health-care setting, a social norms intervention targeted at clinical behaviour change and a RCT). If the study was excluded, the reviewer entered a reason for exclusion. Any disagreements over the recommendation to exclude or the reason for exclusion were flagged up by Covidence and the two reviewers met to discuss. If they were unable to come to a consensus, there was moderation by a third researcher or discussion at a team meeting.

Data extraction and management

For efficiency, data extraction was conducted in three stages: stage 1 involved extraction of all data apart from the details of the intervention, stage 2 was the BCT coding (carried out by a different team concurrently with stage 1) and stage 3 was carried out later, because it relied on data collected during the BCT coding (e.g. we needed to identify which aspects of the intervention were based on social norms to assess the frequency or format of the social norms intervention). Data from included studies were extracted using data extraction forms derived from the Cochrane Effective Practice and Organisation of Care data collection form.³⁹

Stage 1 data extraction

See extraction form in *Appendix 3, Tables 16 and 17*.

Data were independently extracted by two researchers from the data extraction team (MYT, LH, SR and SC). Any disagreements were referred to a third researcher for consideration or discussed at a research team meeting.

Data extracted were:

- setting of the trial (e.g. primary)
- country in which trial was conducted
- design of trial
- aim of the trial
- unit of allocation
- primary outcome
- secondary outcomes
- time points
- statistical analysis
- inclusion/exclusion criteria (and whether or not the inclusion criteria targeted participants based on low target performance)
- methods of recruitment
- number of randomised clusters, if randomised RCT
- subgroups measured (both health workers and patients)
- target behaviour
- total number of patients and health workers randomised
- type of health worker targeted by the intervention
- withdrawals and exclusions (after randomisation)
- number of participants (both patients and health workers) randomised to group
- number of clusters randomised to group, if cluster RCT
- type of control
- outcomes
- quality assessment (risk of bias).

Stage 2 data extraction: coding of behaviour change techniques

See the BCT extraction form in *Appendix 5, Table 19*.

The processes of the third stage of data extraction, along with the accompanied instructions, were refined through piloting (see *Appendix 4* for the final version). Six studies were independently extracted by Sarah Cotterill and Mei Yee Tang. The instructions were refined during this pilot phase and some categories were added to ensure that extraction was as consistent as possible. The piloting process was repeated until a high level of agreement was reached between the two coders.

In the protocol, we had envisaged that we would contact authors for additional information if the data needed to calculate effect sizes were not adequately reported in the paper. We were not able to do this owing to time constraints, but we made efforts to search for additional papers, including process evaluations and protocols. Once all of the data were extracted, they were transferred to Stata for analysis.

Assessment of risk of bias

As part of the first stage of data extraction, risk of bias for each included study was independently assessed by the data extractors (LH, MYT, SC and SR) using The Cochrane Collaboration tool for assessing risk of bias³⁶ across a range of criteria: selection bias, performance bias, detection bias, attrition bias, reporting bias, selective outcome reporting and other biases. Included studies were classified as having a high, low or unclear risk of bias for each criterion. All risk-of-bias criteria were added as part of the data extraction form in Covidence. Where disagreements occurred, a discussion between the two extractors took place to resolve the disagreement or a third data extractor would be brought in when an agreement could not be reached. Percentages of high/low/unclear judgements for each risk-of-bias criterion across included studies were calculated and reported as a bar chart to provide a summary of the risk of bias across criteria domains (see *Figure 2*). Text summaries across each criteria of bias were produced in line with The Cochrane Collaboration's guidance for large reviews.³⁶ Judgements for each risk-of-bias criterion for all included studies were reported (see *Appendix 15, Figure 20*).

Data analysis

Outcomes and prioritisation

The primary outcome for the review was compliance of the health worker with the desired behaviour at the time point closest to 6 months post intervention. We expected studies to report different behaviours (e.g. prescribing, hand-washing and test ordering) and we expected studies to measure those behaviours in different ways. We converted any observed measure of health worker behaviour into a standardised mean difference (SMD) between groups in terms of compliance with the desired behaviour. Common examples included the mean number of times a behaviour was performed per health worker or the mean rate of behaviour (e.g. percentage of the population for whom antibiotic items were dispensed). At times, compliance was reported as a binary outcome, such as compliance versus non-compliance on a single occasion, and was expressed either in a binary format or using an odds ratio.

We used the methods of Chinn⁴⁴ to convert binary outcomes to a SMD with associated standard errors (see *Appendix 8* for the formula).

If several measures of compliance were reported in sufficient detail to enable the analysis of a trial, we used the following criteria to select the outcome for the primary analysis, in decreasing order of importance: (1) observed measure rather than self-report, (2) appropriate adjustment for clustering, (3) continuous measure, (4) final score rather than percentage change or change from baseline, (5) described as the primary outcome, (6) used to calculate the sample size and (7) reported first.

The secondary outcome for the review was patient health-related outcomes that were likely to result from targeting the health worker behaviour. These were converted to a SMD using a similar approach.

Specific BCTs were independently double coded using the BCT taxonomy v1²⁴ by at least two researchers. A BCT extraction form was produced to guide the process. Each study's intervention descriptions from all of the relevant papers (e.g. protocol, process evaluation and main findings) were collated by Mei Yee Tang and transposed to the BCT extraction form so that the second coder (RP, SC or JR) could have the information required for BCT coding available in the one document for each study. All coders had access to the full papers on Covidence, so that they were able to find further relevant information that could help them with the coding task. Following coding, for each study Mei Yee Tang transferred the BCT codes and information extracted by both coders onto a single final BCT extraction form. As part of this process, any discrepancies were highlighted by Mei Yee Tang and the disagreements were resolved through discussion or by the moderation of another coder.

For each study, BCTs were separately coded for all arms (i.e. the control arm and all intervention arms). Mei Yee Tang coded BCTs for all studies, which were then independently double coded by another trained coder within the research team (SC, RP or JR). BCT coders also recorded the target population, target behaviour, whether or not guidelines were provided as part of the intervention and the direction of change in the behaviour that was desired.

Training

All coders completed online training on the coding of BCTs (www.bct-taxonomy.com/; accessed 1 September 2020) and attended a workshop facilitated by co-applicant Rachael Powell and study steering committee member Marie Johnson prior to starting the coding process. To ensure that all coders were familiar with the BCT coding process and coded consistently, a random sample (using www.random.org; accessed 1 September 2020) of three studies was selected for coding by all coders (MYT, RP, SC and JR). All coders coded the three intervention descriptions independently before meeting to discuss any issues that arose. This practice exercise with all BCT coders was repeated again on another four randomly selected studies. The two practice sessions helped to refine the coding process and revise the BCT extraction form (see *Appendix 5, Table 19*). A decision log was kept throughout the BCT coding process to record any decisions that were made to ensure the consistency of coding. Details are provided in *Appendix 6*.

Behaviour change techniques inter-rater reliability

Inter-rater reliability for each of the BCTs that were present at least once across all intervention arms was assessed using the prevalence-adjusted and bias-adjusted kappa (PABAK) statistic (see *Appendix 7, Table 20*), which adjusts for both the prevalence and the occurrence of BCTs.⁴⁰ In circumstances in which prevalence is low, the widely used chance-corrected kappa statistic is likely to underestimate reliability as it is highly dependent on prevalence.⁴¹ To calculate the PABAK, the kappaetc module in Stata® I/C 14.0 (StataCorp LP, College Station, TX, USA) was used to produce the Brennan–Prediger statistic.^{42,43}

Stage 3 data extraction: trial and intervention characteristics

See the stage 3 data extraction form in *Appendix 4, Table 18*.

Information relating to trial and intervention characteristics [focused on the social norms element(s) only] was extracted during stage three of data extraction using Microsoft Excel® (Microsoft Corporation, Redmond, WA, USA) by Sarah Cotterill and Mei Yee Tang:

- Did the inclusion criteria target participants based on low target performance?
- Frequency and intensity of intervention.
- Format of intervention.
- Source of the intervention (i.e. the person delivering the intervention).
- Was this person delivering the intervention internal or external to the target person's organisation?
- Reference group/person used as the comparison/source of approval.
- Type of comparison.

Some trials incorporated baseline measurements into their analyses. This was carried out either by adjusting for baseline values of the outcome measure or of other prognostic variables in an analysis of covariance (ANCOVA), or by reporting outcomes as changes from baseline. We have prioritised ANCOVA-adjusted estimates of the treatment effect where relevant or those from logistic regression, given that these are generally more precise. Change scores cannot be pooled through conversion to SMDs.

Missing data

Our preferred approach to dealing with missing data was to take steps to try to obtain them. We searched for companion papers by author searching and citation searching. Contacting trial authors was not possible owing to limited resources and the large number of studies. We imputed estimates of standard deviations where necessary by using any available information, such as *p*-values, confidence intervals (CIs), ranges or standard errors of baseline data, by pooling standard deviations from other similar studies that use the same type of outcome or by searching for trials that used the same outcome. Where necessary, for cluster randomised trials we imputed a value of the intraclass correlation coefficient (ICC) by pooling across similar studies.

Unit of analysis issues

Where any of the studies in the review were cluster randomised trials, we extracted both raw summary measures (e.g. means and numbers having had the event) and adjusted standard errors from appropriately analysed trials. Where it was not possible to obtain the adjusted SMD and its standard error directly, the methods that were used to calculate the SMD and standard error are shown in *Appendix 9, Table 21*.

Several studies had more than two relevant arms (e.g. two different social norms interventions and a control group). In each case, we extracted data on any comparison that was relevant to our primary research question, while avoiding double counting where possible. Where relevant, we combined study arms that contained identical BCTs. In cases with two different social norms interventions and a single control arm, where possible we divided the number of health-care workers in the control arm approximately evenly between the comparisons to avoid double counting, while retaining the correct intervention effect. In studies with more than one candidate control arm, we chose the comparison that provided the more pure test of social norms (e.g. social norms intervention + X vs. X is a more pure test of social norms than social norms intervention + X vs. usual care).

Where a study was a factorial trial analysed appropriately using linear or logistic regression, we extracted the covariate and standard error that best assessed the effect of social norms BCTs, for example a covariate comparing all arms containing a social norms BCT with all arms without.

Analysis of skewed data

If the primary outcome data were heavily skewed, meta-analyses based on SMDs of the untransformed data would be expected to produce biased estimates. In some cases, compliance was reported as 'mean per cent compliance' or similar, and there is a likelihood that this outcome is skewed when close to 0% or 100% owing to it being bounded. We removed data likely to be skewed (where mean compliance was close to 0% or 100%) in a sensitivity analysis.

Utilising the behaviour change technique coding in the analysis

The approach we took to utilising the BCTs in the meta-analysis was to create an Excel file of all the trials, listing the intervention and control BCTs on separate rows. We subtracted the control arm BCTs from the intervention arm BCTs to identify the BCTs that would be expected to be responsible for the differences between the two arms.

Using the five types of comparison (extracted during the BCT coding process), listed in *Box 1*, allowed us to separate out three different tests of social norms:

1. 'Pure' test of social norms intervention alone (comparisons 1 and 2, see *Box 1*).
This involved trials with social norms BCT(s) in the intervention arm and no BCTs in the control arm (comparison 1). These trials were the purest test of social norms interventions: the BCTs being tested were those found in the intervention arm. For trials in which an intervention arm including a social norms BCT combined with other BCTs was tested against a control arm containing the same other BCTs (comparison type 2), the control arm BCTs were subtracted from the intervention arm to reveal the BCTs that would be expected to account for differences in outcome. For example, if the study tested social comparison and instructions on how to perform the behaviour (intervention arm) against instructions on how to perform the behaviour (control arm), the comparison type would be 'social comparison'.
2. 'Complex' test of a social norms intervention alongside one or more other BCTs (comparison 3, see *Box 1*)
This involved trials in which an intervention arm including a social norms BCT combined with other BCTs was tested against a control arm containing none of the same BCTs (comparison 3). The control arm was deducted from the intervention arm to reveal the test involved in the comparison. For example, if the study tested a complex intervention such as credible source, feedback on behaviour, social support unspecified and behavioural practice/rehearsal (intervention arm) against social support unspecified and behavioural practice/rehearsal (control group), the comparison would be 'credible source' and 'feedback on behaviour' versus control.
3. Social norms intervention occurring in both arms (comparisons 4 and 5; see *Box 1*)
In some studies, two different social norms interventions were compared (comparison 4) or the same social norms intervention appeared in both arms (comparison 5). Where social norms interventions occurred in both arms of a trial, the study did not provide useful information for the meta-analysis, because these trials do not test the effect of social norms interventions, but they were potentially useful to the review as follows:
 - Any study that directly compared one social norms BCT against another (e.g. social comparison vs. credible source) could potentially be included in the network meta-analysis.
 - Any study that compared the same social norms BCT in both arms, with the addition of other BCTs (e.g. with the addition of social support in one of the arms) or comparing differing modes of delivery (e.g. social comparison delivered in person or by e-mail) could potentially be included in the metaregression.
 - Any study where social norms BCT(s) were delivered in both arms as a control intervention, for the purpose of testing a separate intervention, in which the social norm was a minor part were not included in any analysis.

BOX 1 Types of comparison

Comparison

Comparison 1: social norms BCT vs. any control.

Comparison 2: social norms BCT + X vs. X.

Comparison 3: social norms BCT + X vs. any control.

Comparison 4: social norms BCT type A vs. social norms BCT type B.

Comparison 5: social norms BCT + X vs. social norms BCT + Y.

In summary, the information extracted for the analysis describes the BCTs that were tested in the study rather than all of the BCTs that make up the intervention. In some cases (comparison 1) the content of the comparison is the same as the content of the intervention arm, but in most cases (comparison 2 and 3) the content of the comparison is what is left of the intervention when the control arm is taken away. We regard this as the part of the intervention that was actively tested in the trial. A limitation of this approach is that we may have missed some interaction effects.

Feedback on behaviour

Early on in our coding, we observed that the BCT 'feedback on behaviour' was often found to be presented alongside a social norms BCT. The implementation of three social norms BCTs (social comparison, social incentive and social reward) would seem to be greatly facilitated by combination with 'feedback on behaviour'. Social comparison, defined as 'draw attention to others' performance to allow comparison with the person's own performance',²⁴ does not by definition require feedback on the target's own behaviour to be provided, but providing such feedback (e.g. performance data) would be expected to facilitate comparison. Social reward, 'arrange verbal or non-verbal reward if and only if there has been effort and/or progress in performing the behaviour',²⁴ and social incentive, 'inform that a verbal or non-verbal reward will be delivered if and only if there has been effort and/or progress in performing the behaviour',²⁴ similarly do not require feedback on the target's behaviour to be provided (e.g. the behaviour could be monitored by others without feedback to make the reward/incentive process clear to a target), but feedback on the behaviour fits very well with these social norms BCTs and might be expected to facilitate the action of these BCTs.

Because of the high prevalence of feedback on behaviour (present in 88/100 comparisons), we combined 'feedback on behaviour' with the social norms BCT with which it appeared for the purpose of primary analyses: in the forest plots we have listed each social norm with or without feedback. However, it was important to unpick the separate effects of feedback on behaviour: this was examined as part of the metaregression. As a sensitivity analysis, we examined the overall effects of social norms interventions with and without feedback on behaviour.

Data synthesis

Criteria for study data to be meta-analysed

We included in a meta-analysis those studies that report a primary outcome measure (clinical behaviour of a health worker) or secondary outcome (patient outcome) that can be converted into a SMD.

Planned approach for meta-analysis

Research question (RQ) 1: what is the effect of social norms interventions on the clinical behaviour of health workers, and the resulting patient outcomes?

The comparisons used in the analysis to answer RQ1 are shown in *Appendix 10, Table 22*. We stratified the studies in the forest plot according to the type of comparison (see *Utilising the behaviour change technique coding in the analysis*) and the type of target behaviour, and pooled estimates across strata. The aim of this was to provide some initial insight into whether or not, and how, treatment effects vary systematically in trials using different social norms techniques, while remaining aware of the likely confounding by other trial characteristics. We considered I^2 and tau when interpreting heterogeneity, but did not use it as the basis for analytic decisions. We preferred a fixed-effects approach rather than a random-effects approach to meta-analysis, which we consider to yield a summary of the evidence in these trials (i.e. the average effect), rather than an estimate of a common underlying treatment effect. However, we also reported a random-effects analysis.

Research question 2: which contexts, modes of delivery and BCTs are associated with health worker clinical behaviour change? To address this research question, we followed steps 1 to 3.

Step 1 – we explored sources of variation using forest plots and narrative description. In addition to those comparisons used in RQ1, we included the following types of comparison in a narrative description: (1) social norms intervention A versus social norms intervention B and (2) social norms intervention + X versus social norms intervention + Y, where X and Y are any BCT or combination of BCTs, and A and B are either two different types of social norms BCT or the same social norms BCT delivered by two different methods.

Step 2 – we undertook an exploratory analysis using multivariable metaregression to investigate sources of heterogeneity and explain variation in the results. Metaregression is an appropriate regression method in which weights are assigned to studies/subgroups based on the standard error of the treatment effect. *Appendix 11, Table 23*, shows the predictor variables together with anticipated parameterisations that we included in the metaregression analyses. Although controlling for multiple predictors at once is desirable, in practice this was governed by the number of trials and the observed distributions of the variables. We allowed for trials from the different comparisons to enter into a single metaregression given that we anticipated we would be able to control for comparators and co-interventions in the regression. We had intended to categorise the control conditions, but were unable to do this robustly.

Step 3 – we used network meta-analysis to explore which social norms BCT, combination of social norms BCTs or combination of social norms BCT with other BCTs, if any, appears most effective. We considered two broad approaches for network meta-analysis, and made the decision to employ type (a) after consultation with our SSC.

- Network meta-analysis
We examined data from all trials to look at the most commonly occurring combinations of social norms BCTs, either alone or alongside other BCTs. We built and examined a network diagram including social norms BCTs and commonly occurring combinations of social norms BCTs with other BCTs, plus control. Decisions about whether or not to ‘lump together’ BCTs or combinations of BCTs into categories were made after careful discussion by the project team. The justifications were recorded. The geometry of the network diagram was evaluated and no revisions were required to achieve a connected network. Fixed-effects and random-effects network meta-analyses were fitted in Stata.
- Multi-components-based network meta-analysis.⁴⁵
Each intervention in the review would have been considered as a combination of BCT components. We would include all social norms BCTs along with other commonly found BCTs in a components-based network plot. This type of analysis ideally requires all available trials that test the BCT components of interest; our search strategy was not appropriate for this as we were focusing on the social norms components only. We therefore decided not to pursue this approach.

The results from direct and indirect evidence were compared to check for consistency. Trials grouped by comparison were examined to assess transitivity. Metaregression did not identify any clear potential effect modifiers; therefore, although we planned in the protocol to include these in the model, this did not happen.

Additional analyses

We carried out the following sensitivity analyses for the primary outcome:

- include only studies with a low risk of bias (for the key domains of allocation concealment, sequence generation, attrition, selective outcome reporting and other sources of bias)

- exclude continuous outcomes reported as 'mean percentage' that were < 20% or > 80%, as these are unlikely to come from a normal distribution
- include only studies in which the standard deviation was not imputed
- using alternative values of imputed ICC
- studies with and without feedback on behaviour.

Publication bias

We aimed to minimise the impact of reporting biases by performing a comprehensive search for eligible studies. We investigated the impact of publication bias in the reported studies using a funnel plot.

Patient and public involvement

We recruited members of the public from two sources: (1) PRIMER (Primary Care Research in Manchester Engagement Resource: <https://sites.manchester.ac.uk/primer/about-primer/>; accessed 1 September 2020), a public involvement group in the Centre for Primary Care, University of Manchester, and (2) an advertisement on Citizen Scientist, which is based at Salford Royal NHS Foundation Trust and promotes research and patient and public involvement (PPI) opportunities for members of the public. We advertised for anyone aged > 18 years who had used any type of NHS service: we were not looking for people with any particular condition or experiences.

Mr Manoj Mistry has been involved in the review from the start. He has a wealth of past experience of involvement in research and was an invaluable part of the review. He had input into the proposal before we submitted the funding bid and he was a member of the SSC, bringing a patient and carer perspective to the meetings. He attended all three SSCs and played a full and active role in the committee's discussions.

Two PPI events were planned for this study. The first event took place in August 2018 at the University of Manchester. The aim of the first event was to discuss how the review would be relevant to members of the public, and to get feedback on the overall design of the review. Six members of the public (two female, four male), including Mr Manoj Mistry, participated in the workshop, and they discussed with us the relevance of the review for patients and carers. They felt that patients can have a role in changing health worker behaviour, for example by reminding health workers to wash their hands or telling the GP that they do not expect to be prescribed antibiotics for a cold, although they were cynical about whether or not doctors would listen to patients when they present potential best practice (example given of a relative who had better care in Australia, but the doctor in the UK did not want to hear about it). In response to this observation, we made sure to record whether or not any studies in the review considered patients' role in social norms interventions (e.g. use of the information about others' approval BCT, where the approval came from patients) (see *Appendix 12, Table 24*, for a short report of the meeting).

We had feedback from four public contributors on the *Plain English summary*, and Mr Manoj Mistry has reviewed this description and account of our PPI activity.

A second PPI workshop took place in October 2019 at the University of Manchester to discuss how best to disseminate the findings to a wider audience. Four of the original group members (including Mr Manoj Mistry) attended. We presented the preliminary findings from the SOCIAL study and asked the group what they considered to be the most important messages from a public perspective. We also asked the group to suggest suitable language for presenting the findings to a lay audience. They suggested that the main messages should be that the study provides evidence that social norms interventions can encourage the medical community to change behaviour, leading to better outcomes for patients.

METHODS

One or two things make social norms interventions even more effective:

- right message (i.e. the use of different social norms BCTs)
- right place (i.e. context)
- right method (i.e. mode of delivery).

Authority of the message sender is crucial.

Messages from all sources are important, including those from patients.

We plan to follow this approach when we write summary materials for a lay audience. There was concern (from some) about the term 'behaviour change'. Alternatives were 'influence' or 'improve', but they did not all agree. The group wanted us to avoid being preachy or patronising or using a telling-off approach to health workers: they talked about health workers being 'encouraged' by social norms interventions, rather than 'directed'. The group felt that social norms messages would also be useful with people who teach and mentor students and young professionals. The lack of effect for face-to-face delivery of social norms interventions was viewed as surprising.

TABLE 1 Characteristics of included studies

Study characteristic (<i>n</i> = 106)	Frequency	%
Country		
Australia	8	7.5
Canada	15	14.2
Denmark	4	3.8
UK	13	12.3
Netherlands	6	5.7
USA	45	42.5
Other/multiple	15	14.2
Setting		
Primary (GP/general practice nurses)	57	53.8
Hospital (inpatient and outpatient)	31	29.3
Community	4	3.8
Care/nursing home	4	3.8
Mixed	7	6.6
Other	3	2.8
Type of health worker		
Doctor (primary care)	45	42.5
Doctor (secondary care)	19	17.9
Other (nurse/dentist/AHP/pharmacist)	7	6.6
Mixture/whole team	35	33.0
Target behaviour		
Prescribing (including vaccinations)	40	37.7
Hand-washing/hygiene	4	3.8
Tests/assessments	21	19.8
Referrals	3	2.8
Management communications	25	23.6
Other	2	1.9
Multiple behaviours	11	10.4
Type of trial		
Cluster RCT	69	65.1
Factorial	4	3.8
RCT	28	26.4
Stepped wedge	4	3.8
Matched pairs, cluster RCT	1	0.9
Targeted at low baseline performance? ^a		
No	103	97.2
Yes	2	1.9
Unclear	1	0.9

Chapter 3 Results

Identification of included studies

Of the 7980 studies identified using database searches, 3552 were identified as duplicates leaving 4428 separate studies to be screened. Of these, 3951 were discarded as irrelevant to the research questions under consideration, leaving 477 to be assessed for eligibility by full-text review of publications. Of these, 361 were excluded as ineligible for various reasons, as described in *Figure 1*. There were 116 studies that met the inclusion criteria, and 106 of these contributed findings to the review. Some of the 106 studies had more than one trial arm, and there were a total of 117 comparisons that tested the effect of social norms on the clinical behaviour of health workers. The remaining 10 studies met all of the inclusion criteria but did not provide usable outcome data: two reported the overall effect but did not compare the results between groups,^{46,47} six reported results unclearly or incompletely,⁴⁸⁻⁵³ one trial was discontinued before completion⁵⁴ and one did not report results on our primary or secondary outcomes.⁵⁵ Searches for companion papers were unsuccessful and authors were not contacted owing to limited time. A brief description of the studies is provided in *Appendix 13, Table 25*.

Characteristics of included studies

Study characteristics

A detailed summary of study characteristics is shown in *Table 1*. Over half of the included trials were conducted in North America (Canada: $n = 15$, 14.2%; USA: $n = 45$, 42.5%) and the most common settings were primary care ($n = 57$, 53.8%) and hospitals (including both inpatient and outpatient: $n = 31$, 29.3%). GPs were the most frequently targeted type of health worker ($n = 45$, 42.5%), with

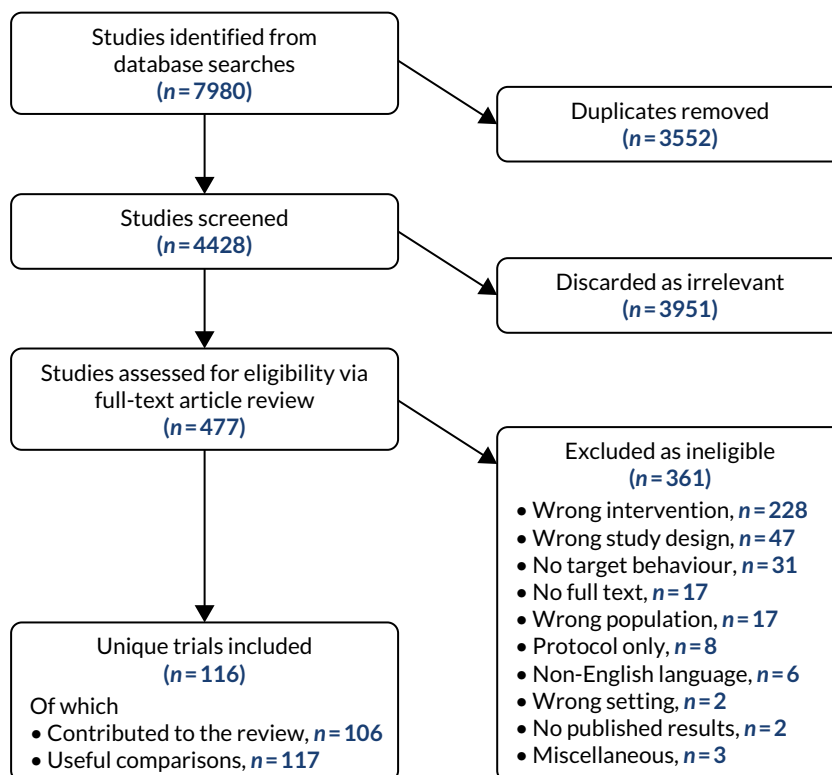


FIGURE 1 The PRISMA flow chart of the SOCIAL review.

many studies also targeting a mixture of health workers ($n = 35$, 33.0%). In terms of target behaviour, 40 studies (37.7%) aimed to change prescribing behaviours (including vaccinations), with 25 (23.6%) concerned with the overall management of conditions/communications (e.g. being friendly during consultations) and 21 (19.8%) focusing on arranging, conducting or administering tests/assessments (e.g. performing HbA_{1c} testing). Of the 106 trials that contributed findings to the review, the majority ($n = 70$, 66%) were cluster RCTs, with 31 RCTs (29.2%), four stepped-wedge designs (3.8%) and one two-arm matched-cluster RCT. The majority of trials ($n = 103$, 97.2%) did not explicitly target participants with a low target performance. This is surprising, because the literature on social norms strongly suggests that social comparison is more likely to be successful if it is addressed to low performers: telling a high performer that they are already doing more than their peers does not motivate them to improve.^{56,57} It is possible that some of the trials took place in contexts where the performance of all the health professionals was generally low at baseline, so they did not need to specifically seek out low performers to target, but no information was provided to support such an assumption. A complete list of all trials and their characteristics is included in *Appendix 14, Table 26*.

Intervention characteristics

Details of intervention characteristics are shown in *Table 1*. Of the 117 comparisons, many were delivered using a written (paper) format ($n = 29$, 24.8%) or utilised a mixed format (e.g. face to face and written) ($n = 18$, 15.4%). Participants in 45 (38.5%) interventions received the intervention more than twice, whereas 10 comparisons delivered the intervention twice (8.5%) and 35 (29.9%) delivered the intervention only once. In the majority of comparisons, the investigators were the source of the intervention ($k = 83$, 70.9%) and the intervention was delivered by someone external to the target health worker's organisation ($k = 81$, 69.2%). In 97 (82.9%) of the comparisons, the reference group/person was the target health worker's peer. In terms of the desired direction of change, 85 (72.6%) studies aimed to increase the behaviour. There was a lack of clarity in reporting across many intervention characteristics within the included studies. For example, in 34 (29.1%) interventions, the format was unclear or not reported and the frequency of the intervention was unclear or not reported in 27 (23.1%) comparisons.

Description of the behaviour change techniques

The frequency of specific social norms BCTs occurring within the 100 comparisons that tested social norms interventions against a control are shown in *Table 2*. We found tests of social comparison ($n = 79$), credible source ($n = 7$) and social reward ($n = 2$) against control. Some studies tested more than one social norms BCT together: social comparison and credible source ($n = 6$), social comparison and social reward ($n = 2$) and multiple social norms BCTs (more than two) together ($n = 4$). The social norms interventions often occurred alongside other BCTs, and 22 different techniques were identified (see *Table 2*).

TABLE 2 Frequency of behaviour change techniques occurring in comparisons

Behaviour change technique	<i>n</i>
Social norms BCTs	
6.2 Social comparison	90
9.1 Credible source	18
10.4 Social reward	5
6.3 Information about others' approval	4
10.5 Social incentive	1
Other BCTs	
2.2 Feedback on behaviour	88
3.1 Social support (unspecified)	25

TABLE 2 Frequency of behaviour change techniques occurring in comparisons (*continued*)

Behaviour change technique	n
4.1 Instruction on how to perform the behaviour	20
7.1 Prompts/cues	19
5.1 Information about health consequences	18
1.2 Problem-solving	12
1.1 Goal-setting (behaviour)	9
8.1 Behavioural practice/rehearsal	5
1.4 Action planning	4
6.1 Demonstration of the behaviour	4
1.3 Goal-setting (outcome)	3
2.3 Self-monitoring of behaviour	2
2.7 Feedback on outcome(s) of behaviour	2
12.5 Adding objects to the environment	2
1.7 Review outcome goals	1
2.1 Monitoring of behaviour by others without feedback	1
5.3 Information about social and environmental consequences	1
8.2 Behaviour substitution	1
9.2 Pros and cons, final	1
10.3 Non-specific reward, final	1
12.1 Restructuring the physical environment	1
10.1 Material incentive (behaviour)	1

The table totals more than 100 because some interventions included multiple BCTs. There are 100 eligible comparisons.

The 117 comparisons belonged to the three comparison categories as follows (*Table 3*):

- Pure comparisons. There were 36 comparisons that offered a 'pure' test of social norms interventions. Most of these tested social comparison (33 comparisons). There were far fewer comparisons testing credible source ($n = 3$), social reward ($n = 1$), social comparison and credible source together ($n = 2$), or social comparison and social reward together ($n = 2$).
- Comparisons involving other BCTs. Social comparison was combined with social support (unspecified) ($n = 7$), prompts and cues ($n = 5$), information about health consequences ($n = 4$) and instruction on how to perform the behaviour and prompts/cues (combined) ($n = 5$). Combined interventions involving social comparison with more than two other BCTs or where the combination occurred only once in the study were combined into one group: social comparison and other BCTs ($n = 25$). Credible source did not occur more than once with any one particular BCT, so there is one category of credible source with other BCTs ($n = 4$) and another of social comparison and credible source with other BCTs ($n = 4$). There was one example of social reward with other BCTs ($n = 1$). Where more than two social norms BCTs occurred together, they were combined in a category ($n = 4$).
- Social norms BCTs in both arms. There were 17 comparisons in which both arms involved social norms interventions (*Table 4*).

TABLE 3 Social norms comparison types

Social norms interventions comparison types	n (%)
Pure comparisons	
Social comparison	33 (28)
Credible source	3 (3)
Social reward	1 (1)
Social comparison and credible source	2 (2)
Social comparison and social reward	2 (2)
Comparisons involving other BCTs	
Social comparison and social support (unspecified)	7 (6)
Social comparison and prompts/cues	5 (4)
Social comparison and information about on health consequences	4 (3)
Social comparison and instruction on how to perform the behaviour & prompts/cues	5 (4)
Social comparison and other BCTs	23 (21)
Credible source and other BCTs	4 (3)
Social comparison and credible source and other BCTs	6 (3)
Social reward and other BCTs	1 (1)
Multiple social norms BCTs and other BCTs	4 (3)
Social norms BCTs in both arms	
Social norms BCTs both arms	17 (15)
Total (n)	117

TABLE 4 Comparisons that involve social norms interventions in both arms

Comparison type	Frequency
Same social norms intervention in both arms, testing some other intervention (n = 11)	
Credible source and feedback on behaviour vs. credible source and feedback on behaviour with other BCTs	1
Social comparison vs. social comparison and goal-setting	1
Social comparison and feedback vs. social comparison and feedback plus information about others' approval and other BCTs	1
Social comparison and feedback vs. social comparison and feedback with other BCTs	7
Social comparison and feedback vs. social comparison and feedback with a patient-level intervention	1
Comparison of two social norms interventions (n = 2)	
Credible source and social comparison and feedback on behaviour and other BCTs vs. social comparison and feedback on behaviour	1
Social comparison and credible source and feedback on behaviour vs. credible source	1
Testing different variants of the same social norms intervention (n = 4)	
Social comparison vs. social comparison, no other BCTs	2
Credible source vs. credible source	1
Social comparison and feedback and social support (unspecified) vs. social comparison and Feedback plus social support (unspecified)	1
Total	17

Variation in trial characteristics by type of social norms intervention

Table 5 shows the key trial characteristics by the type of comparison. In total, 33 different comparisons were a pure test of 'social comparison' and these were quite varied in terms of type of target behaviour, type of health-care worker and type of setting; similarly, those trials that tested social comparison alongside other BCTs were also quite varied. There were 13 comparisons that tested 'credible source' alone or with other BCTs and four that included social reward, and again these were spread over a range of behaviours, contexts and settings. Reassuringly, there is no clear pattern to suggest that the use of BCTs was restricted to particular behaviours, contexts or settings, and this is consistent with the regression results, which suggested that the results were consistent after adjustment.

Outcome measures

Using the criteria described in *Chapter 2, Outcomes and prioritisation*, we selected a single primary outcome measure of compliance with desired behaviour for each relevant comparison. Of the 117 comparisons used in the review, 32 (27%) provided an odds ratio, 42 (35%) provided raw binary data and 43 (37%) provided mean with standard deviation (standard deviations were imputed where necessary).

Risk of bias

A summary of each risk-of-bias item across the included studies ($n = 106$) is shown in *Figure 2*. Individual risk-of-bias assessments for all of the included studies can be found in *Appendix 15, Figure 20*.

Allocation

In terms of random sequence generation, methods were deemed sufficient to produce comparable groups for the majority of studies ($n = 78$, 73.6%) and were, therefore, considered to be at low risk of bias. Only one study was rated to be at high risk of bias, and this was because of the original randomisation being rejected because of a perceived lack of balance between groups. All other studies ($n = 27$) were rated as unclear because of insufficient information to permit a judgement. The majority of studies were also rated to be at low risk of bias for allocation concealment ($n = 78$, 73.6%), primarily because of recruitment and consent being conducted before randomisation took place. Three studies were rated to be at high risk of bias because randomisation took place before recruitment and/or obtaining consent. The remaining studies ($n = 25$) were considered unclear because there was insufficient information available, or because recruitment/consent had occurred post randomisation and it was unclear whether or not participants were aware of their allocation at the time of enrolment/consent.

Blinding

Many of the studies were cluster trials, randomised at the hospital or clinic level, making the blinding of participants and personnel impractical. Owing to this clustering, most studies were rated to be at high risk of bias ($n = 85$). Fourteen studies were considered to be at low risk of bias owing to participants/personnel being unaware of the study aims and hypothesis, intervention content and existence of other groups/interventions, or being unaware that they were taking part in a trial. Studies rated as unclear ($n = 7$) lacked clarity in reporting the blinding of participants and whether or not participants were aware of the intervention or study aims and outcomes. For blinding of outcome assessment, most studies ($n = 80$) obtained data from electronic health records, online reports or databases or routinely collected data, in which the outcome assessors were blind to group allocations and were, therefore, rated to be at low risk of bias. Eight studies were judged to be at high risk of bias where participants selected consecutive patient records to contribute to the outcome assessment (so could have selected groups based on good practice), where participants selected the patients and collected the data or

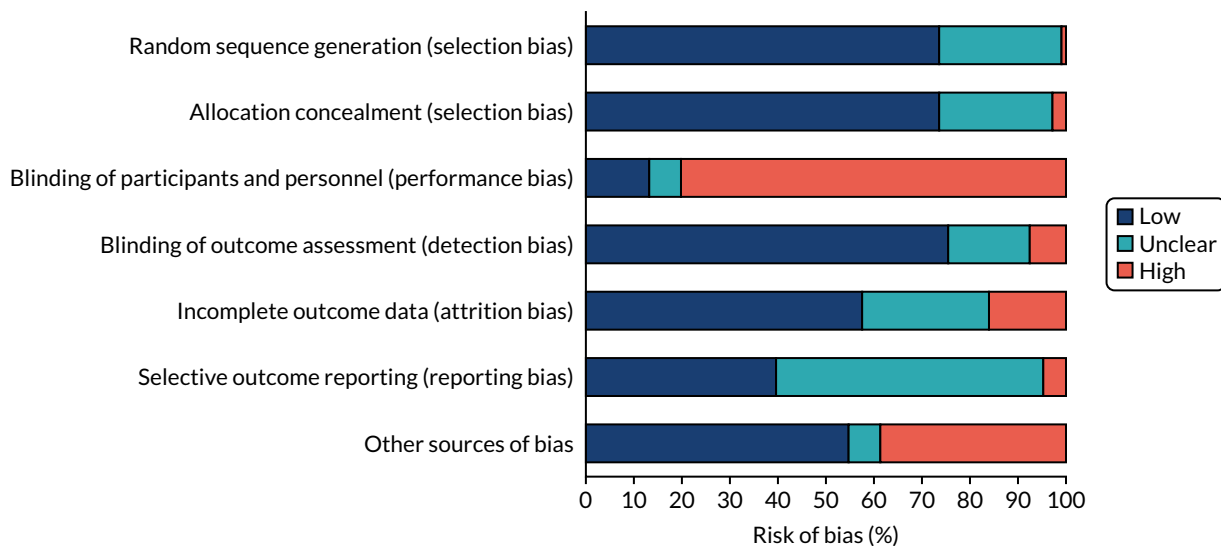


FIGURE 2 Review authors' judgements about each risk-of-bias item presented as percentages across all included studies ($n = 106$).

where outcome measures were recorded by participants. The remaining studies ($n = 18$) were judged to be unclear because of a lack of information or lack of clarity in terms of whether or not outcome assessors were blinded to allocation.

Incomplete outcome data

Over half of the included studies ($n = 61$) were considered to be at low risk of bias for incomplete outcome data for the following reasons: no drop-outs/low attrition, attrition evenly distributed across groups, drop-out reasons unlikely to be connected to interventions, analysis conducted on intention-to-treat basis and different patients before and after. Seventeen studies were rated to be at high risk of bias as they had large numbers of drop-outs, a lack of discussion of drop-out reasons, unequal (or unreported) attrition across groups or reasons for drop-outs being related to the intervention. The remaining studies ($n = 28$) were judged to be unclear as a result of insufficient or unclear reporting of attrition.

Selective reporting

Where outcomes appeared to have been reported as stated in the protocol (or where there was only one outcome and this was adequately reported), studies were judged to be at low risk of bias for selective outcome reporting ($n = 42$). Five studies had made changes from the planned outcomes in the protocol, or had ambiguously reported the primary outcome at registration (enabling multiple interpretations), and were, therefore, considered to be at high risk of bias. Studies that had no available protocol and had multiple outcomes, or several ways of reporting outcomes, were categorised as being unclear ($n = 59$).

Other potential sources of bias

No potential other sources of bias were identified for over half of the included studies ($n = 58$). A small number of studies ($n = 7$) were rated as unclear for reasons including a lack of clarity in whether or not adjustments were made for clustering in the analysis, an unclear unit of analysis for some statistical tests and only summary trial methods and data reported. The remaining studies ($n = 41$) were judged to be at high risk of bias for the following reasons: no adjustment for clustering in analyses; poor reporting (particularly of outcome data, therefore making interpretation difficult); potential problems with the study design (e.g. stepped-wedge, step-wise regression); concerns over analysis processes (e.g. extreme values replaced or excluded); important baseline differences between groups or large differences in participant numbers across arms; and analysis within rather than between groups.

TABLE 5 Key trial characteristics by social norm comparison type

Trial characteristic	Social norm comparison category, n (%)													
	Social comparison source	Credible source	Social reward	Social comparison and credible source	Social comparison and social reward	Social comparison and social support (unspecified)	Social comparison and prompts and cues	Social comparison and information on health consequences	Social comparison and instructions and prompts/cues	Social comparison and others BCTs	Credible source and other BCTs	Social comparison and credible source and other BCTs	Social reward and other BCTs	Multiple social norms and other BCTs
Comparisons with primary outcome data (n)	33	3	1	2	2	7	5	4	5	25	4	4	1	4
Target behaviour														
Prescribing	15 (45)		1 (100)	1 (50)	2 (100)	2 (29)	1 (20)	3 (75)	1 (20)	11 (44)	1 (25)	1 (25)		1 (25)
Hand/hygiene												1 (25)	1 (100)	1 (25)
Tests	7 (21)					1 (14)	3 (60)	1 (25)	3 (60)	4 (16)	1 (25)	1 (25)		
Referrals										2 (8)		1 (25)		
Manage conditions	5 (15)	3 (100)		1 (50)		2 (29)			1 (20)	5 (20)	2 (50)			2 (50)
Other						12 (14)				1 (4)				
Multiple	6 (18)					1 (14)	1 (20)			2 (8)				
Type of HCP														
Doctor: GP	16 (48)			1 (50)	2 (100)	4 (57)	2 (40)	1 (25)	4 (80)	11 (44)	2 (50)	1 (25)		1 (25)
Doctor: secondary	4 (12)	3 (100)				1 (14)	1 (20)	1 (20)		2 (12)	1 (25)	1 (25)		1 (25)
Other HCP	4 (12)		1 (100)							1 (4)	0 (0)			1 (25)
Mixed/team	9 (27)			1 (50)		2 (29)	2 (40)	2 (40)	1 (20)	10 (40)	2 (50)	2 (50)	1 (100)	1 (25)
Setting														
Primary	18 (55)			1 (50)	2 (100)	4 (57)	4 (80)	1 (25)	5 (100)	17 (68)	2 (50)	1 (25)		1 (25)
Hospital	6 (18)	3 (100)				2 (29)	1 (20)	2 (50)		6 (24)	2 (50)	2 (50)	1 (100)	2 (50)
Community	1 (3)		1 (100)	1 (50)										1 (25)
Care/nursing	0 (0)					1 (14)		1 (25)		1 (4)		1 (25)		
Mixed	7 (21)													
Other	1 (3)													

HCP, health-care professional.

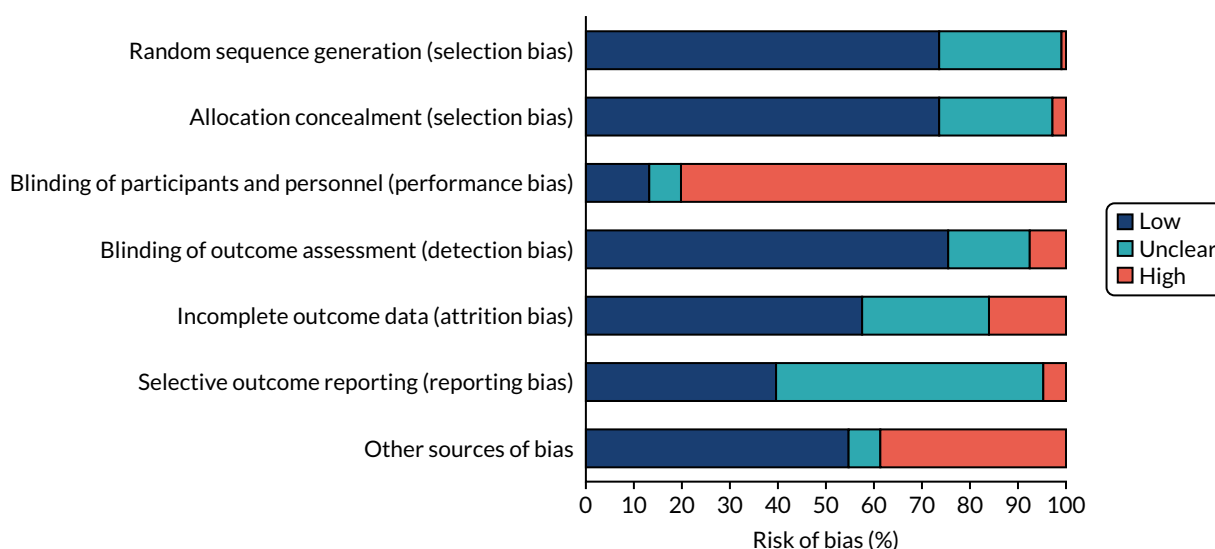


FIGURE 2 Review authors' judgements about each risk-of-bias item presented as percentages across all included studies ($n = 106$).

where outcome measures were recorded by participants. The remaining studies ($n = 18$) were judged to be unclear because of a lack of information or lack of clarity in terms of whether or not outcome assessors were blinded to allocation.

Incomplete outcome data

Over half of the included studies ($n = 61$) were considered to be at low risk of bias for incomplete outcome data for the following reasons: no drop-outs/low attrition, attrition evenly distributed across groups, drop-out reasons unlikely to be connected to interventions, analysis conducted on intention-to-treat basis and different patients before and after. Seventeen studies were rated to be at high risk of bias as they had large numbers of drop-outs, a lack of discussion of drop-out reasons, unequal (or unreported) attrition across groups or reasons for drop-outs being related to the intervention. The remaining studies ($n = 28$) were judged to be unclear as a result of insufficient or unclear reporting of attrition.

Selective reporting

Where outcomes appeared to have been reported as stated in the protocol (or where there was only one outcome and this was adequately reported), studies were judged to be at low risk of bias for selective outcome reporting ($n = 42$). Five studies had made changes from the planned outcomes in the protocol, or had ambiguously reported the primary outcome at registration (enabling multiple interpretations), and were, therefore, considered to be at high risk of bias. Studies that had no available protocol and had multiple outcomes, or several ways of reporting outcomes, were categorised as being unclear ($n = 59$).

Other potential sources of bias

No potential other sources of bias were identified for over half of the included studies ($n = 58$). A small number of studies ($n = 7$) were rated as unclear for reasons including a lack of clarity in whether or not adjustments were made for clustering in the analysis, an unclear unit of analysis for some statistical tests and only summary trial methods and data reported. The remaining studies ($n = 41$) were judged to be at high risk of bias for the following reasons: no adjustment for clustering in analyses; poor reporting (particularly of outcome data, therefore making interpretation difficult); potential problems with the study design (e.g. stepped-wedge, step-wise regression); concerns over analysis processes (e.g. extreme values replaced or excluded); important baseline differences between groups or large differences in participant numbers across arms; and analysis within rather than between groups.

Effects of interventions: health worker behaviour (primary outcome)

Overall effect

There were 100 comparisons suitable for meta-analysis. Figure 3 shows the SMD summarised by type of BCT comparison in a fixed-effects meta-analysis. From this plot we can see that, as expected, there is a large amount of heterogeneity with an overall I^2 -value of 85.4%. Overall, combined data suggest that, on average, interventions that include social norms components were associated with a modest improvement of 0.08 SMD (95% CI 0.07 to 0.10). Forest plots showing each individual study in the review, by the social norm BCT used in the study, are in Appendix 17, Figures 22–26.

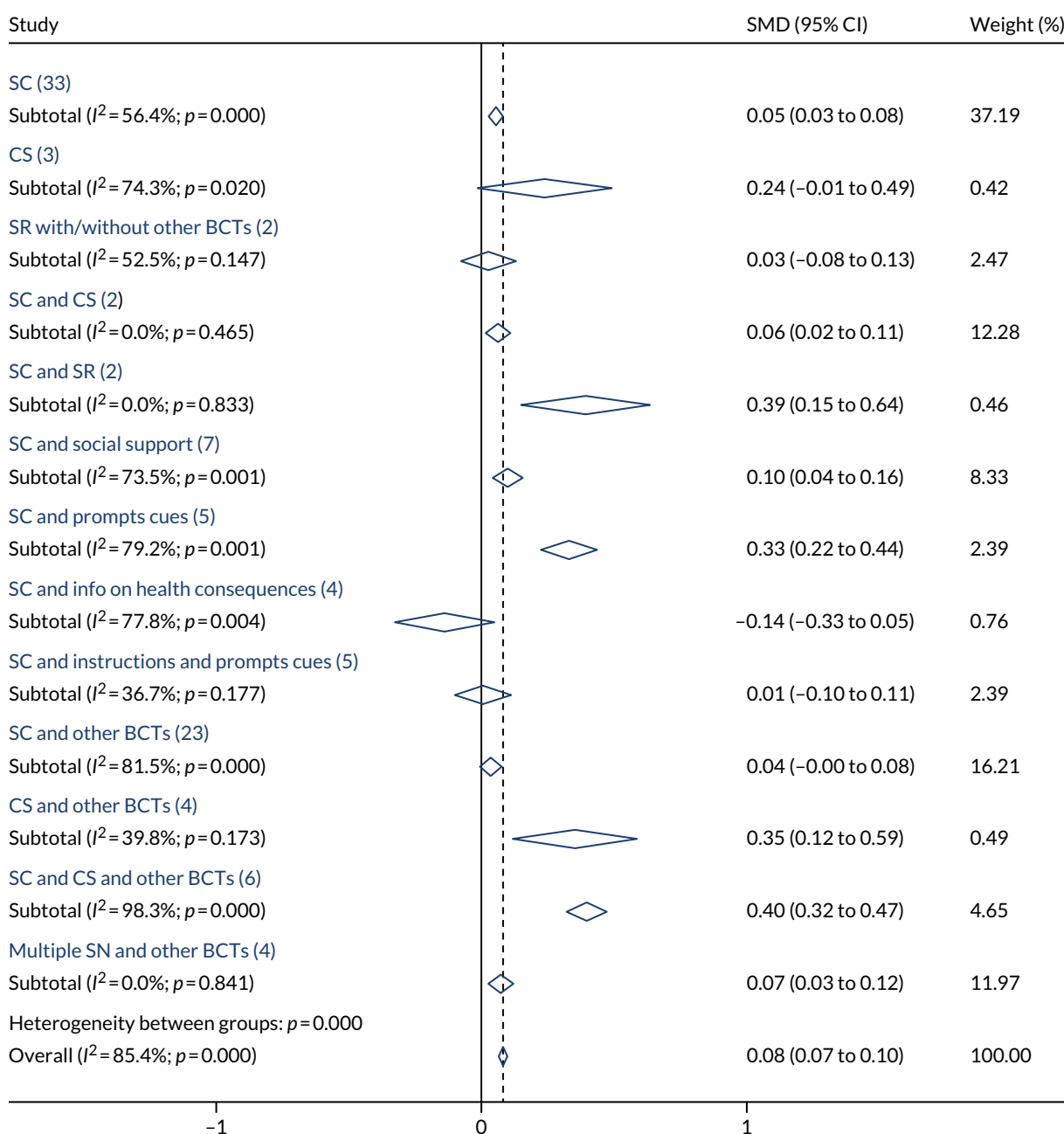


FIGURE 3 Fixed-effects forest plot summarised by type of comparison. CS, credible source; SC, social comparison; SR, social reward. Note that SR (one comparison) and SR and other BCTs (one comparison) have been combined in this graph to improve presentation.

The I^2 is interpreted as a measure of the proportion of variability owing to heterogeneity between studies. This can be calculated only when two or more studies are included in the same subgroup/meta-analysis, as can the p -value alongside I^2 that tests the null hypothesis that $I^2 = 0$. Note that I^2 is related to precision and rapidly approaches 100% when the number of studies is large.⁵⁸ τ^2 is an alternative measure of heterogeneity, calculated only during a random-effects meta-analysis, and can be interpreted as the between-study variance.

Note that most trials in this review were randomised at a cluster level and the unit of analysis may be patient, health-care worker or a larger unit, such as clinic or hospital. With that in mind, it is impossible to report 'N' for each trial in any consistent way. In the meta-analysis, the weights were calculated based on the standard error of the SMD extracted from the individual trials, which has been adjusted for clustering where necessary. The number of comparisons is reported for each subgroup in brackets on all forest plots. Forest plots showing the effect in every individual study, summarised by social norms intervention, are included in *Appendix 1* (see *Figures 22–26*).

Figure 4 shows the SMD, summarised by the type of BCT comparison in a random-effects meta-analysis. Similar conclusions can be drawn when looking at the random-effects result compared with the fixed-effects result; however, using weights from a random-effects meta-analysis suggests a larger overall SMD (0.16, 95% CI 0.11 to 0.21, $I^2 = 85.4\%$, $\tau^2 = 0.043$) and a wider CI, because the random-effects meta-analysis attributes more weight to smaller trials.

Illustration of standardised mean differences

For this review, we converted all measures of the effectiveness of health worker behaviour into a common scale: the SMD (standardised effect size). SMDs can be difficult to interpret. To illustrate how the observed average standardised effect sizes translate into real health-care scenarios, we have converted the SMD into a risk difference (difference in percentage points) for a range of typical baseline compliance rates (*Table 6*). This was carried out in two steps: (1) transforming the SMD into an odds ratio using a method suggested in the Cochrane handbook³⁶ section 12.6.3, and (2) transforming the odds ratio into a risk difference, using a method proposed by Grant.⁵⁹

Investigation of social norms behaviour change techniques

Note that owing to the high prevalence of the BCT feedback on behaviour (present in 88/100 comparisons), in the forest plots we have combined feedback on behaviour with the social norms BCT with which it appeared; that is we have listed each social norms BCT with or without feedback. Later we examine the separate effect of feedback on behaviour as part of the metaregression and as a sensitivity analysis.

We summarised the SMDs by the type of social norms comparison (see *Figure 3*). There is little consistency in SMDs when looking at the different types of social norms interventions being tested, with subgroup CIs that do not overlap each other and that are inconsistent with the overall effect. Interventions including credible source appear to have larger effect sizes on average than other social norms interventions, and this is true for both credible source on its own ($n = 3$) (SMD 0.24, 95% CI -0.01 to 0.49) and credible source combined with other BCTs ($n = 4$) (SMD 0.35, 95% CI 0.12 to 0.59), although the CIs are wide. Credible source combined with social comparison ($n = 2$) had an average effect of 0.06 SMD (95% CI 0.02 to 0.11), and credible source combined with social comparison and various other BCTs ($n = 6$) appeared to be the most effective of all, with a SMD of 0.40 (95% CI 0.32 to 0.47). Comparisons that were a 'pure' test of social comparison (with or without feedback, $n = 33$) appeared to have a small effect (SMD 0.05, 95% CI 0.03 to 0.08), and the size of the effect is similar when social comparison is combined with various other BCTs ($n = 23$) (SMD 0.04, 95% CI -0.00 to 0.08). Social comparison appeared to be very effective when combined with prompts/cues ($n = 5$) (SMD 0.33, 95% CI 0.22 to 0.44), but ineffective when combined with both prompts/cues and instruction on how to perform the behaviour ($n = 5$) (SMD 0.01, 95% CI -0.10 to 0.11). Social reward appeared to be very effective when combined with social comparison ($n = 2$) (SMD 0.39, 95% CI 0.15 to 0.64), but only

RESULTS

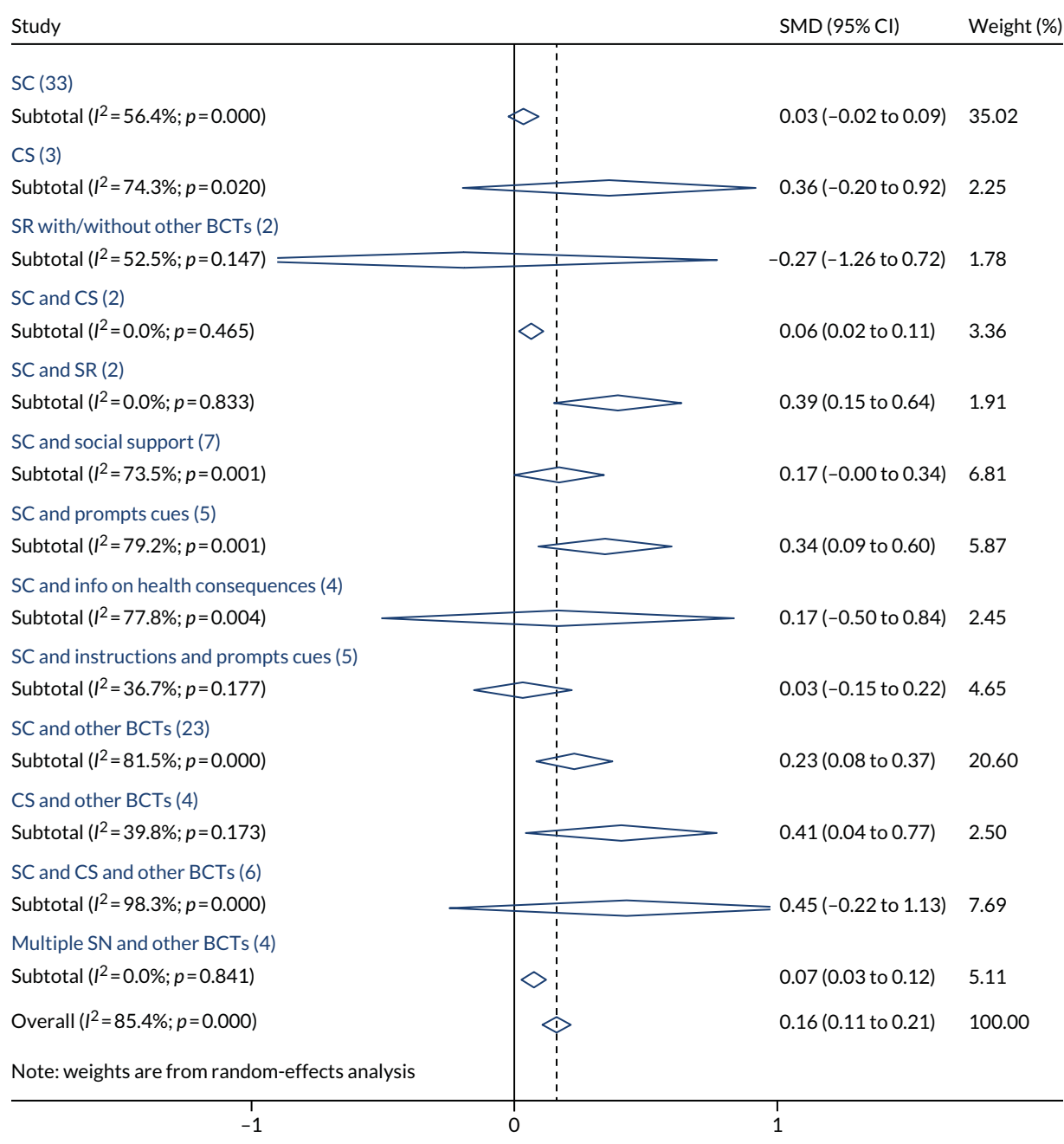


FIGURE 4 Random-effects forest plot summarised by type of comparison. CS, credible source; SC, social comparison; SR, social reward.

one study looked at social reward on its own and found a negative effect. We need to interpret these observations cautiously owing to the large amount of heterogeneity and the differences in contexts and settings.

In an attempt to ease interpretation, *Figure 5* shows a re-categorisation of *Figure 3*. In this plot, all comparisons that test each of the social norms BCTs (social comparison, credible source and social reward), whether alone or alongside other BCTs, have been combined. Trials that combine two or more social norms BCTs have been put together in one group. As before, comparisons that test credible source ($n=7$), either alone or in combination with other BCTs, appear to be the most effective on average (SMD 0.30, 95% CI 0.13 to 0.47). The effect of social comparison ($n=77$) appears to be very small (SMD 0.06, 95% CI 0.04 to 0.8). There is little evidence to suggest that social reward is effective

TABLE 6 Illustration of SMDs

Typical baseline compliance	Expected improvement (percentage points)		Illustration
	Social norms interventions on average 0.08 SMD	Credible source interventions on average 0.3 SMD	
20%	2	10	In a population in which the appropriate tests are ordered 20% of the time, we would expect a social norms intervention, on average, to increase the rate of compliance by 2 percentage points to 22%
40%	4	13	In a population in which prescribing guidelines are being adhered to 40% of the time, we would expect a credible source intervention, on average, to increase the rate of compliance by 13 percentage points to 53%
60%	3	12	In a population in which recommended referrals are being made 60% of the time, we would expect a social norms intervention, on average, to increase the rate of referral by 3 percentage points to 63%
80%	2	10	In a population in which the rate of antibiotic prescribing is 80%, we would expect a credible source intervention, on average, to reduce the rate of prescribing by 10 percentage points to 70%

Note that these values were chosen for illustrative purposes only.

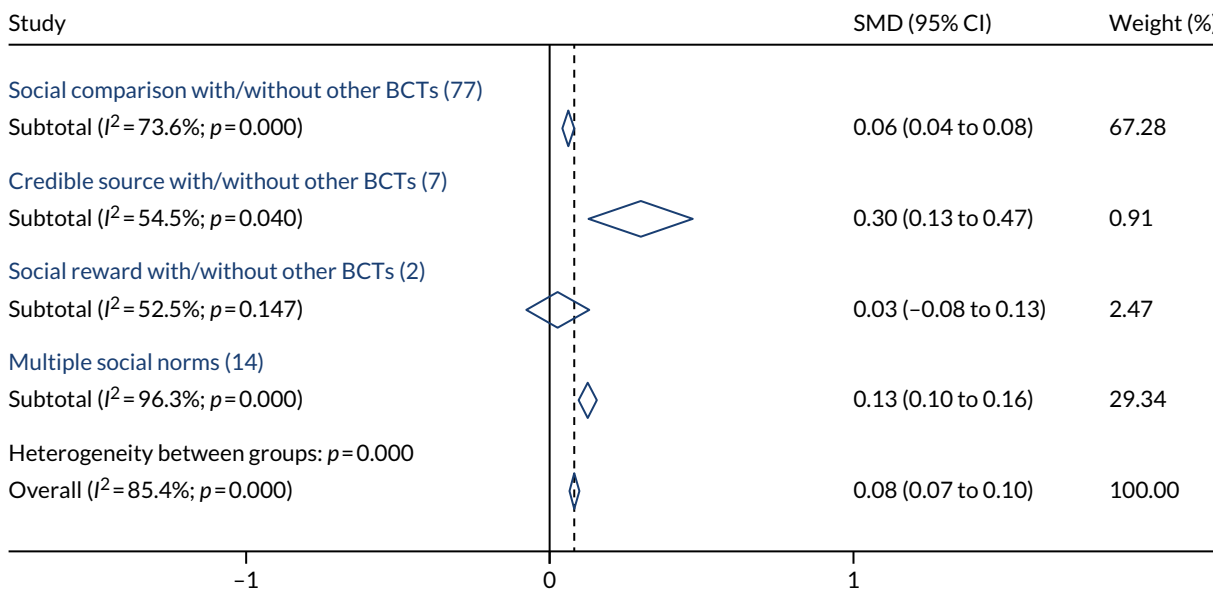


FIGURE 5 Fixed-effects forest plot summarised by alternative categorisation of BCTs.

(SMD 0.03, 95% CI -0.08 to 0.13), but this is based on only two trials. Trials involving a mixture of more than one social norms BCT ($n = 14$) have a larger than average effect (SMD 0.13, 95% CI 0.10 to 0.16).

Illustrative case studies

The purpose of this review is to offer a structured summary across all of the 106 studies, but we have included some illustrative case studies (*Table 7*) to provide a concrete example of each of the three intervention types that were found to be the most effective (credible source and social comparison, social comparison with prompts/cues, social comparison and social reward).

Variation in context and mode of delivery

We have summarised the SMD in various contexts and modes of delivery to examine where, how and with whom social norms interventions are most likely to be effective.

Type of health-care worker

Figure 6 shows the SMD summarised by the type of health-care worker in a fixed-effects meta-analysis. The effect of social norms interventions appears to be quite consistent when comparing GPs with doctors in secondary care as the type of health-care worker targeted (not shown), with an overall effect with doctors ($n = 68$) of 0.08 SMD (95% CI 0.07 to 0.10). We found no evidence that social norms interventions were effective with nurses or allied health professionals (AHPs) (SMD -0.01, 95% CI -0.12 to 0.11), although the number of comparisons was small ($n = 5$). The effect with other health workers, many of which were mixed groups such as doctors/nurses or nurses/AHPs ($n = 27$), was 0.08 SMD (95% CI 0.07 to 0.10).

Target behaviour

Figure 7 shows the SMD summarised by the type of target behaviour in a fixed-effects meta-analysis. Interventions targeting prescribing behaviour ($n = 40$) appeared to be the most effective, on average (SMD 0.11, 95% CI 0.09 to 0.13). The effect of social norms interventions appeared to be reasonably consistent across other types of target behaviour, including test ordering ($n = 21$) and management of/communication about conditions ($n = 23$). Social norms interventions appear to be less effective with hand-washing ($n = 3$, SMD 0.04, 95% CI -0.05 to 0.13) and referrals ($n = 3$, SMD -0.08, 95% CI -0.23 to 0.07), but the number of studies is small.

We have not presented a forest plot summarising the SMD by whether or not the participants were targeted based on low baseline performance, because there were only two trials that did this.

Health-care setting

Figure 8 shows the SMD summarised by health-care setting in a fixed-effects meta-analysis. The effect of social norms interventions appeared to be slightly lower in primary care settings (SMD 0.07, 95% CI 0.05 to 0.09) ($n = 56$) than in hospital settings (SMD 0.12, 95% CI 0.07 to 0.18) ($n = 27$), but both are consistent with the overall effect. Trials taking place in community settings ($n = 4$) and care/nursing home settings ($n = 4$) appear to be less effective on average; however, both CIs do overlap with the overall effect. Trials conducted in mixed settings ($n = 9$) appear to be more effective on average (SMD 0.35, 95% CI 0.27 to 0.42).

Reference group

The reference group is the person or persons that the target is compared with or receives approval from. *Figure 9* shows the SMD summarised by the type of reference group within the trials. The effect of social norms interventions appeared to be reasonably consistent across different types of reference group, with most CIs overlapping, and there was general consistency of each group with the overall effect. Most trials ($n = 84$, 82.9%) had peers as the reference group (SMD 0.08, 95% CI 0.06 to 0.10). Only one trial had patients as the reference group; the effect was consistent with other studies (SMD 0.10, 95% CI -0.17 to 0.37) but the CI was wide because of the low weight in the review.

TABLE 7 Case studies: summary descriptions of interventions, for example studies of the three intervention types found to be the most effective

Details of study	Outcome measure, SMD (95% CI)	Control arm	Intervention description (BCTs coded)
Credible source and social comparison			
Hallsworth <i>et al.</i> (2016) ⁶⁰	The rate of antibiotic items dispensed per 1000 population	Delayed intervention (after the end of the trial)	A letter was sent to GPs from the Chief Medical Officer. The letter stated that the practice was prescribing antibiotics at a rate higher than 80% of practices in its NHS local area team, and used three concepts from the behavioural sciences. The first was social norms information about how the recipient's practice's prescribing rate compared with other practices in the local area. Second, the letter was addressed from a high-profile figure, with the assumption that this would increase the credibility of its content. Finally, the letter presented three specific, feasible actions that the recipient could do to reduce unnecessary prescriptions of antibiotics: giving patients advice on self-care, offering a delayed prescription and talking about the issue with other prescribers in his or her practice. The letter was accompanied by a copy of the patient-focused 'Treating your infection' leaflet, which acted to reinforce the message of the letter by supporting delayed or reduced prescribing
RCT	0.13 (0.03 to 0.29)	No BCTs were coded	
Doctor (primary care)			
Aim: to reduce the number of unnecessary prescriptions of antibiotics by GPs in England			
Social comparison and prompts/cues			
Vellinga <i>et al.</i> (2016) ⁶¹	Adherence to guidelines for antimicrobial prescribing in primary care	Phase 1: a coding workshop – routine coding for UTIs using standardised codes were demonstrated. The purpose of this was to facilitate the generation of electronic A&F reports (not available to control until after the trial). Control practices then provided 'usual care' for the remainder of the intervention	Arm A: phase 1 – a coding workshop (same as control). Phase 2: interactive workshops were designed to promote changes in antimicrobial prescribing for the treatment of UTIs by presenting an overview of prescribing and antimicrobial resistance, discussing the role of the GP in the spread of AMR. A computer prompt was developed for use within the selected general practice management software system. This prompt summarised the recommendations for first-line antimicrobial treatment and appeared on the computer screen when the GP entered the International Classification of Primary Care code (U71) for 'cystitis, urinary infection, other'. This prompt also reminded the GP to collect patients' mobile telephone numbers. Electronic A&F reports were available to download by GPs. These reports provided the practice with information on antimicrobial prescribing for UTIs in comparison with the aggregated information from the other practices participating in the intervention
Arm A			
Cluster RCT	0.55 (0.32 to 0.77)	No BCTs were coded	
Doctor (GP)			
Aim: to increase the number of first-line antimicrobial prescriptions for suspected UTIs in adult patients			
(9.1 credible source, 6.2 social comparison, 2.2 feedback on behaviour, 4.1 instruction on how to perform the behaviour)			
(7.1 prompts/cues, 2.2 feedback on behaviour, 6.2 social comparison)			

continued

TABLE 7 Case studies: summary descriptions of interventions, for example studies of the three intervention types found to be the most effective (continued)

Details of study	Outcome measure, SMD (95% CI)	Control arm	Intervention description (BCTs coded)
Social comparison and social reward			
<p>Persell <i>et al.</i> (2016)⁶²</p> <p>2 × 2 × 2 factorial</p> <p>Doctor (GP)</p> <p>Aim: to reduce inappropriate antibiotic prescribing for ARIs</p>	<p>Physician rate of oral antibiotic prescribing for non-antibiotic-appropriate ARIs, acute sinusitis/pharyngitis and all other diagnoses of respiratory infection</p> <p>0.44 (−0.06 to 0.94)</p>	<p>Intervention 1 (accountable justifications): clinicians received electronic health record alerts summarising the treatment guidelines corresponding to the ARI diagnosis for which the antibiotic was being written, prompted the clinician to enter a free-text justification for prescribing an antibiotic, and informed the clinician that the free-text justification provided would be included in the patient's medical record in which it would be visible to other clinicians. Clinicians were also informed that if no free-text justification was entered, a default statement 'No justification for prescribing antibiotics was given' would appear in the record. If the antibiotic order was cancelled, no justification was required, and no default text appeared. Alerts were suppressed for patients with comorbid chronic conditions that exempted these patients from clinical guidelines (4.1 instruction on how to perform the behaviour, 7.1 prompts/cues). Intervention 2 (suggested alternatives): when entering an ARI diagnosis for a patient, clinicians received a computerised alert containing multiple non-antibiotic prescription and non-prescription medication choices as well as educational materials that could be printed and given to the patient (7.1 prompts/cues)</p>	<p>Intervention 3 (peer comparison): clinicians received e-mailed monthly performance feedback reports that included the clinician's individual antibiotic prescribing rates for non-antibiotic-appropriate ARIs and, as a benchmark, the antibiotic prescribing rate for clinicians who were in the 10th percentile within the clinic (i.e. the lowest rates of inappropriate antibiotic prescribing). If clinicians were among the 10% of their peers with the lowest prescribing rates, the e-mailed reports told clinicians 'You are a top performer.' If clinicians were not among the 10% best, the e-mailed report told clinicians 'You are not a top performer. You are prescribing too many unnecessary antibiotics.' The proportion of 'Top Performers' could be > 10% of clinicians if > 10% of clinicians had an inappropriate antibiotic prescribing rate of zero</p> <p>(2.2 feedback on behaviour, 6.2 social comparison, 10.4 social reward)</p>
AMR, antimicrobial resistance; ARI, acute respiratory infection; UTI, urinary tract infection.			

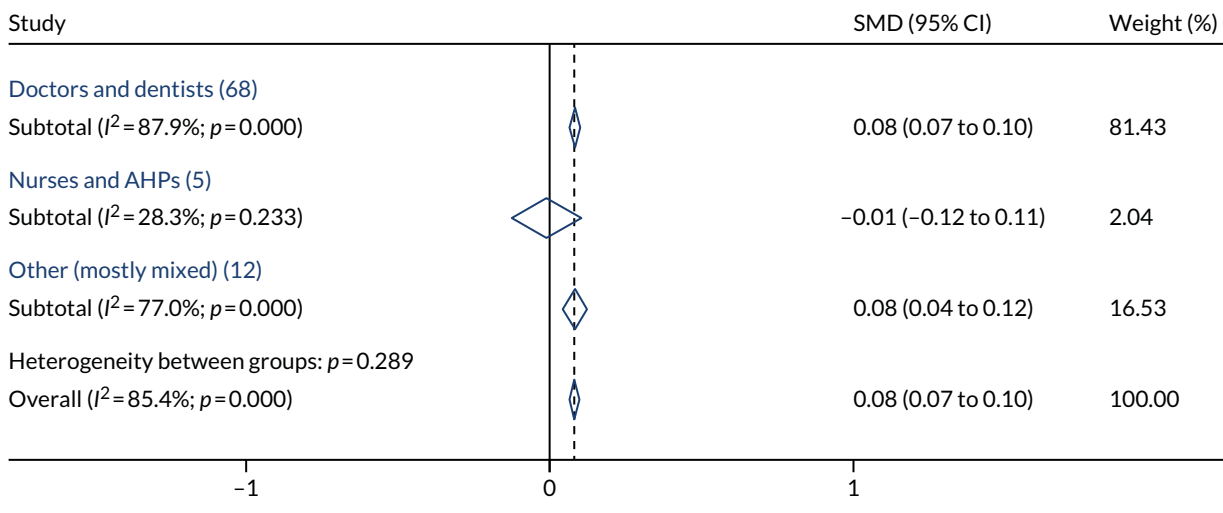


FIGURE 6 Fixed-effects forest plot summarised by type of health-care worker.

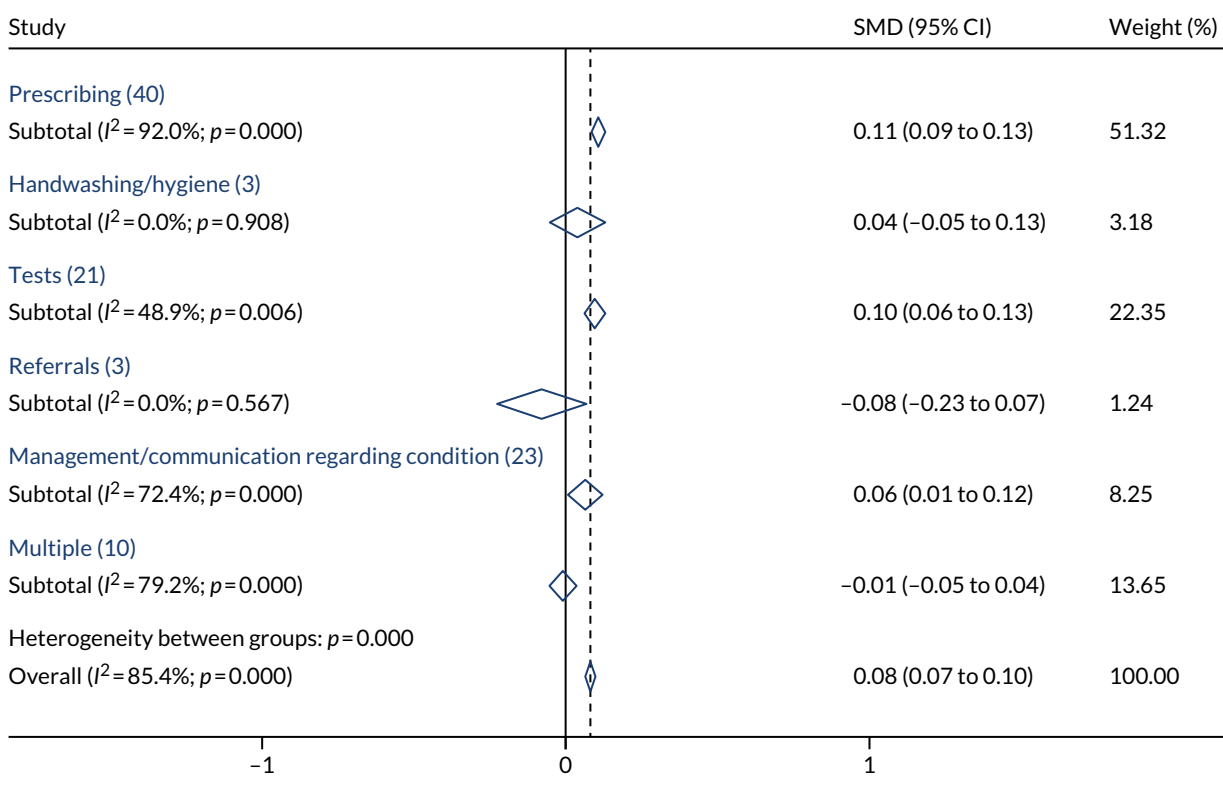


FIGURE 7 Fixed-effects forest plot summarised by target behaviour.

RESULTS

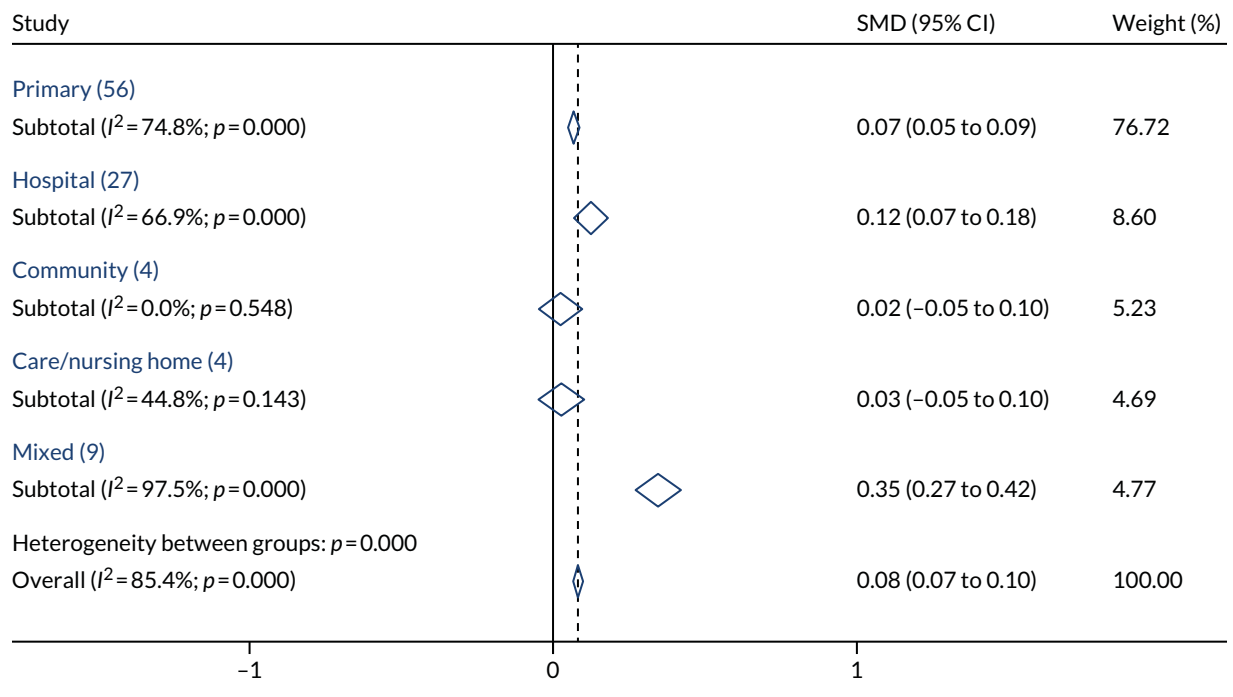


FIGURE 8 Fixed-effects forest plot summarised by health-care setting.

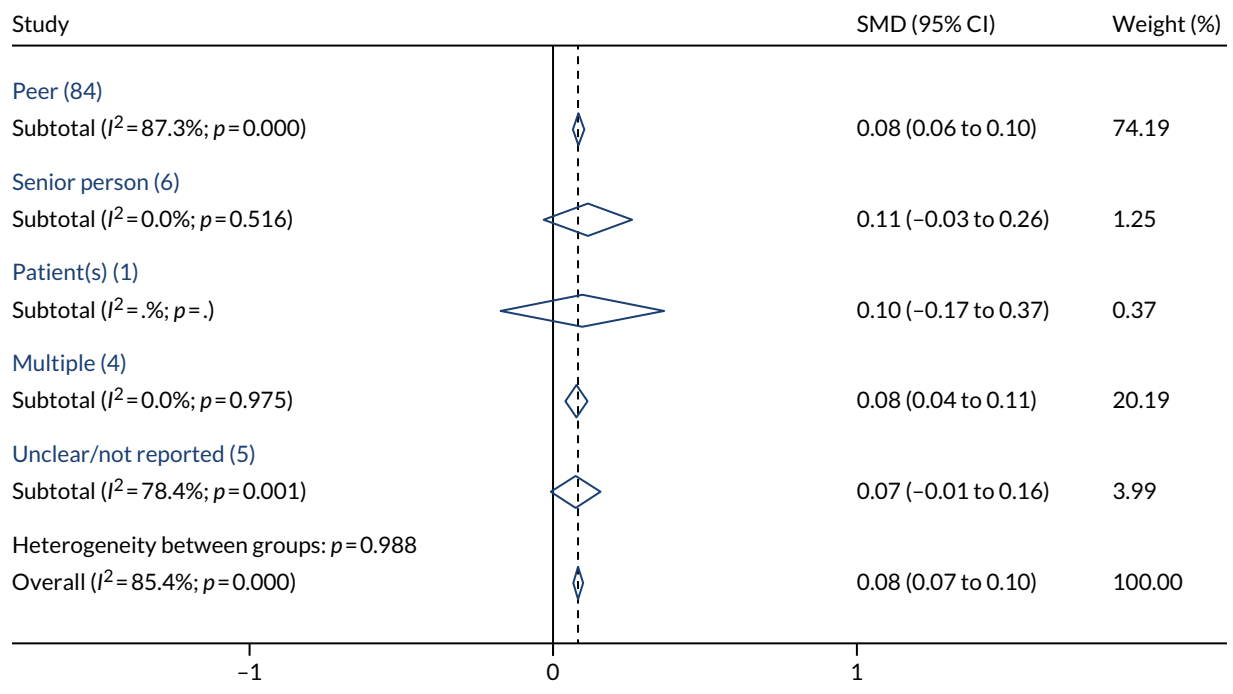


FIGURE 9 Fixed-effects forest plot summarised by reference group.

Benchmarks

When social comparison interventions are delivered, they sometimes include a benchmark: this may be a peer-related benchmark, such as the top 10% or 20% of performers among their peer group, or it may be an external benchmark, such as a performance target set by a royal college. If no benchmark is set, the social comparison usually reports the average performance among peers. The downside of the average approach is that the above-average performers will receive feedback suggesting that they are already performing better than their peers, which may lead them to reduce their effort.⁵⁶ *Figure 10* shows the SMD summarised by the type of benchmark that was used. Only trials involving social comparison have been included, because benchmarking is not relevant to the other social norms interventions. The effect of social norms interventions appeared to be reasonably consistent, regardless of whether a peer benchmark (13 studies: SMD 0.06, 95% CI 0.02 to 0.11) or the average performance (67 studies: SMD 0.11, 95% CI 0.09 to 0.13) was included: the CIs overlap and there is general consistency of each group with the overall effect.

Source of the intervention

Figure 11 shows the SMD summarised by the source of the intervention (i.e. the person delivering the intervention) in a fixed-effects meta-analysis. The effect of social norms interventions appeared to be consistent across the different sources, with the exception of supervisor/senior colleague ($n = 2$), which appeared to be, on average, less effective (SMD -0.28 , 95% CI -0.56 to 0.01). In most trials, the source of the intervention was the investigator ($n = 72$) (SMD 0.08 , 95% CI 0.06 to 0.10) or a respected source ($n = 11$) (SMD 0.09 , 95% CI 0.03 to 0.16). The credible sources that were found in the literature included:

- nurses in management positions who encouraged change in various behaviours to improve hospital stroke care⁶³
- a 'highly respected senior clinician' who persuaded doctors of the harms and limited diagnostic benefit of X-ray for lower back pain⁶⁴
- maternal–fetal medicine specialists, perinatologists or obstetricians who were influential with colleagues, who championed the use of corticosteroids to colleagues in antenatal care⁶⁵
- nurse facilitators with master's degrees and specialist training who promoted changes in preventative care in general practices⁶⁶
- opinion leaders nominated by a peer for their expertise in obstetric care⁶⁷ or breast cancer surgery⁶⁸
- a clinical co-ordinator regarded as a 'credible role model' in managing patients with congestive heart failure⁶⁹
- a letter to poorly performing GPs from the Chief Medical Officer about their rates of antibiotic prescribing.⁶⁰

The other categories occurred infrequently and we should interpret these results with caution owing to some wide CIs.

Direction of change targeted

Figure 12 shows the SMD summarised by the intended direction of change in the behaviour in a fixed-effects meta-analysis. The social norms intervention appeared to be, on average, slightly less effective when the intervention was aimed at increasing a behaviour ($n = 70$) (e.g. more hand-washing) (SMD 0.06 , 95% CI 0.04 to 0.09) than when it was aimed at decreasing a behaviour ($n = 28$) (e.g. prescription of antibiotics) (SMD 0.10 , 95% CI 0.08 to 0.12), but both CIs are consistent with the overall effect.

Frequency of the intervention

Figure 13 shows the SMD summarised by the frequency/intensity of the interventions in a fixed-effects meta-analysis. The effect of the social norms interventions appeared to be, on average, most effective when the intervention was delivered only once ($n = 28$) (SMD 0.25 , 95% CI 0.21 to 0.30). It appeared to be less effective, on average, when delivered more frequently ($n = 47$) (SMD 0.06 , 95% CI 0.04 to 0.08).

RESULTS

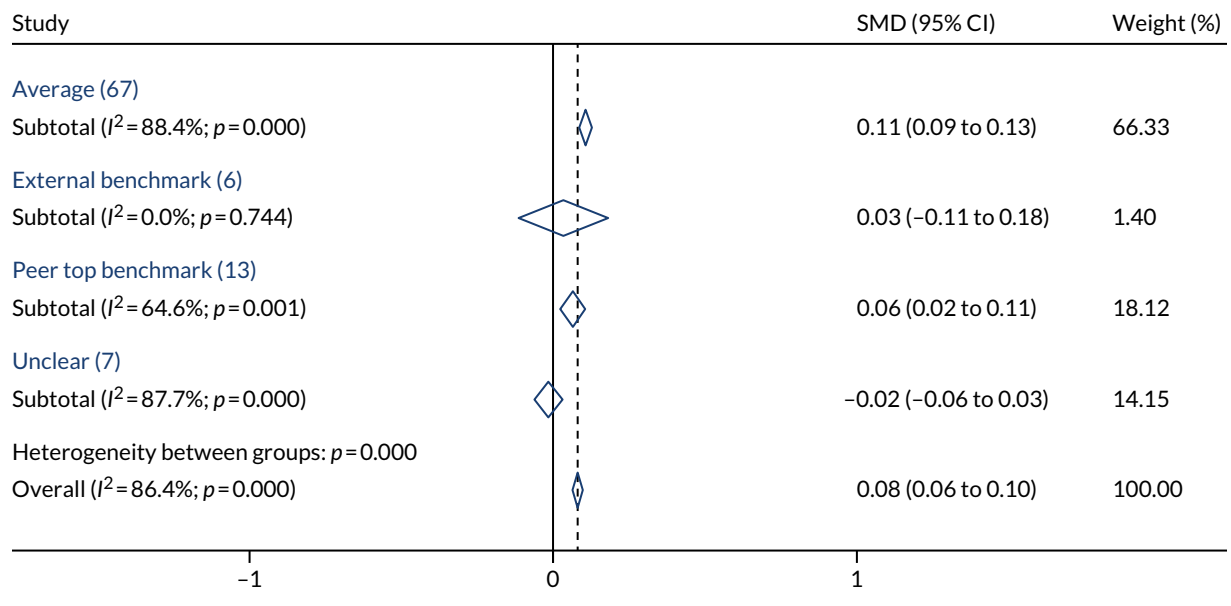


FIGURE 10 Fixed-effects forest plot summarised by benchmark.

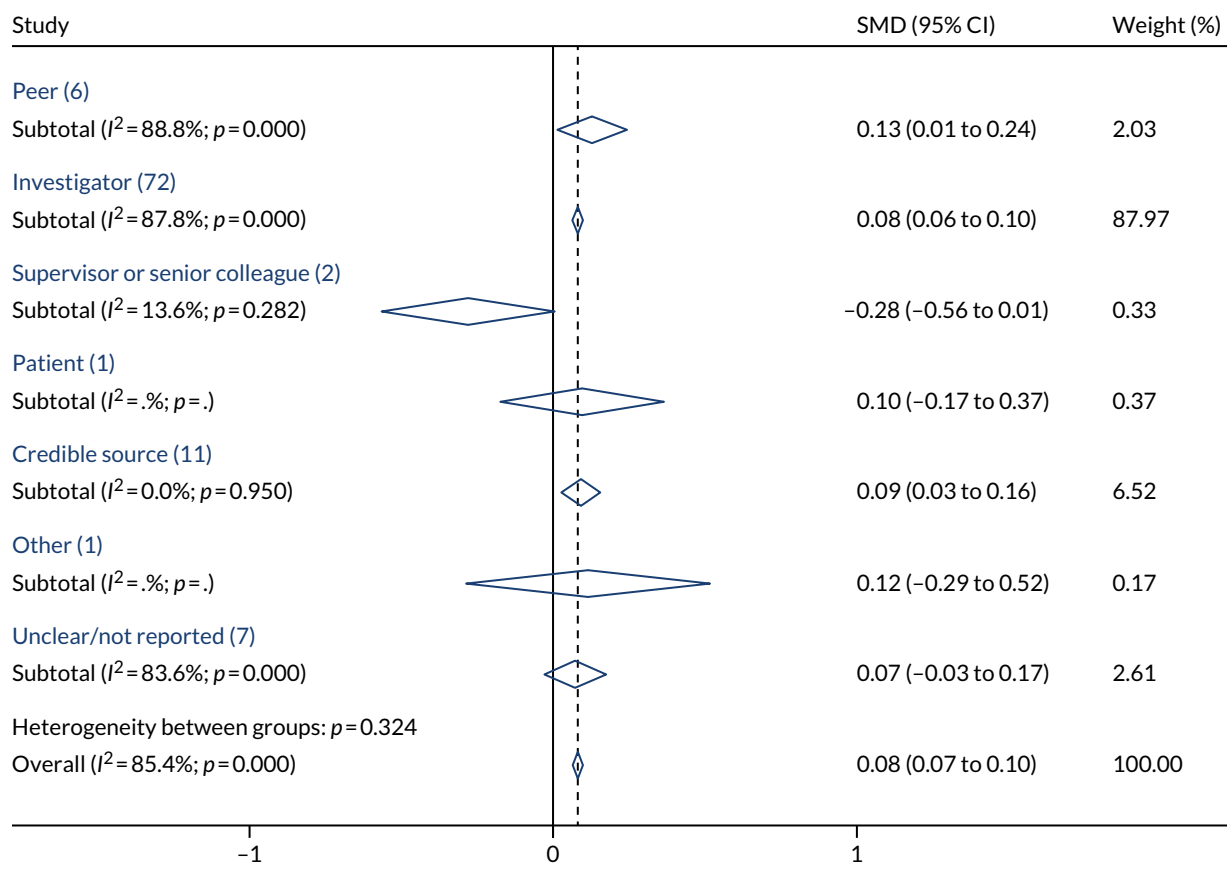


FIGURE 11 Fixed-effects forest plot summarised by source of intervention.

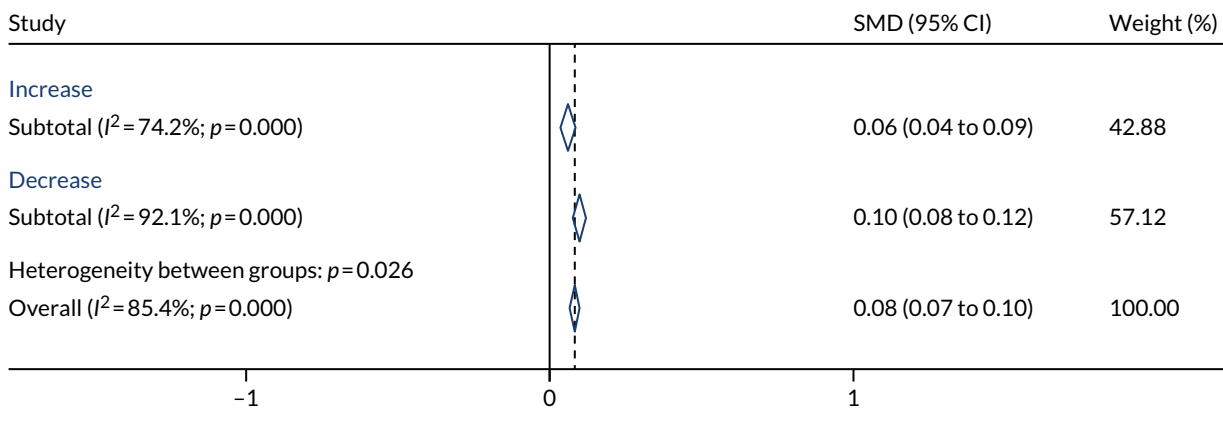


FIGURE 12 Fixed-effects forest plot summarised by the direction of change targeted.

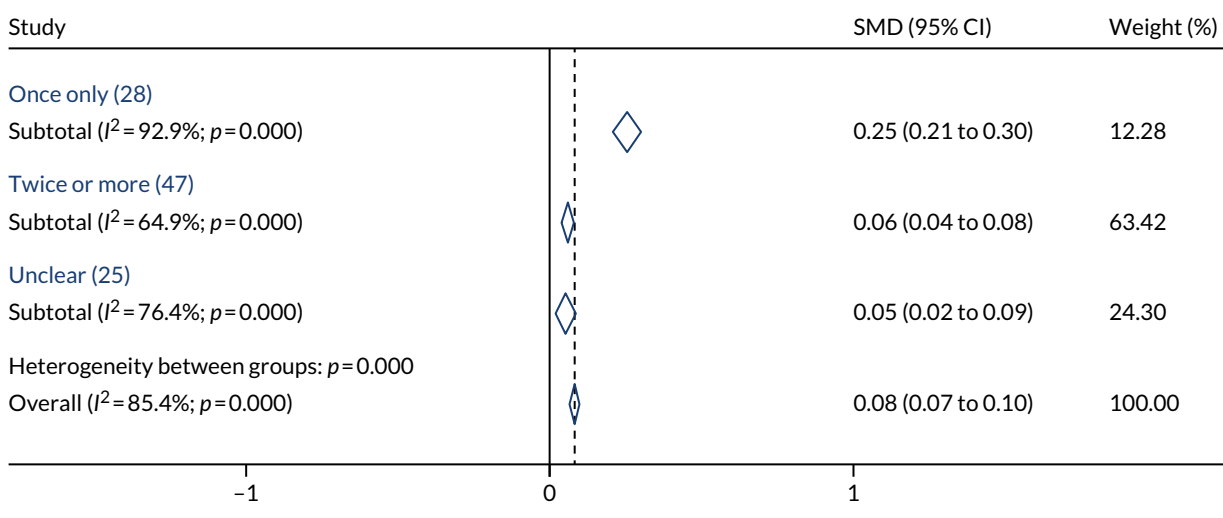


FIGURE 13 Fixed-effects forest plot summarised by frequency of the intervention.

Format of the intervention

Figure 14 shows the SMD summarised by the format of the intervention. Trials delivered via computerised methods whereby the intervention was posted on a website or other computerised format that was not integrated into the health-care worker's workflow ($n = 8$) appeared to be more effective than average (SMD 0.23, 95% CI 0.15 to 0.31). By contrast, interventions delivered face to face ($n = 14$) appeared to be ineffective, on average, with a SMD of -0.01 (95% CI -0.06 to 0.03). Trials with an e-mailed ($n = 9$), written ($n = 25$) or a mixed format ($n = 14$) appeared to be reasonably consistent with each other and with the overall effect.

Person delivering the intervention

Figure 15 shows the SMD summarised by whether the person who delivered the intervention was internal or external to the target's organisation in a fixed-effects meta-analysis. Most ($n = 68$) interventions were delivered by an external person, often the investigator. The effect of social norms interventions, on average, seemed to be consistent across internal sources ($n = 17$) (SMD 0.11, 95% CI 0.05 to 0.17) and external sources ($n = 68$) (SMD 0.08, 95% CI 0.06 to 0.10). However, this should be interpreted cautiously given the wide CIs for internal sources.

RESULTS

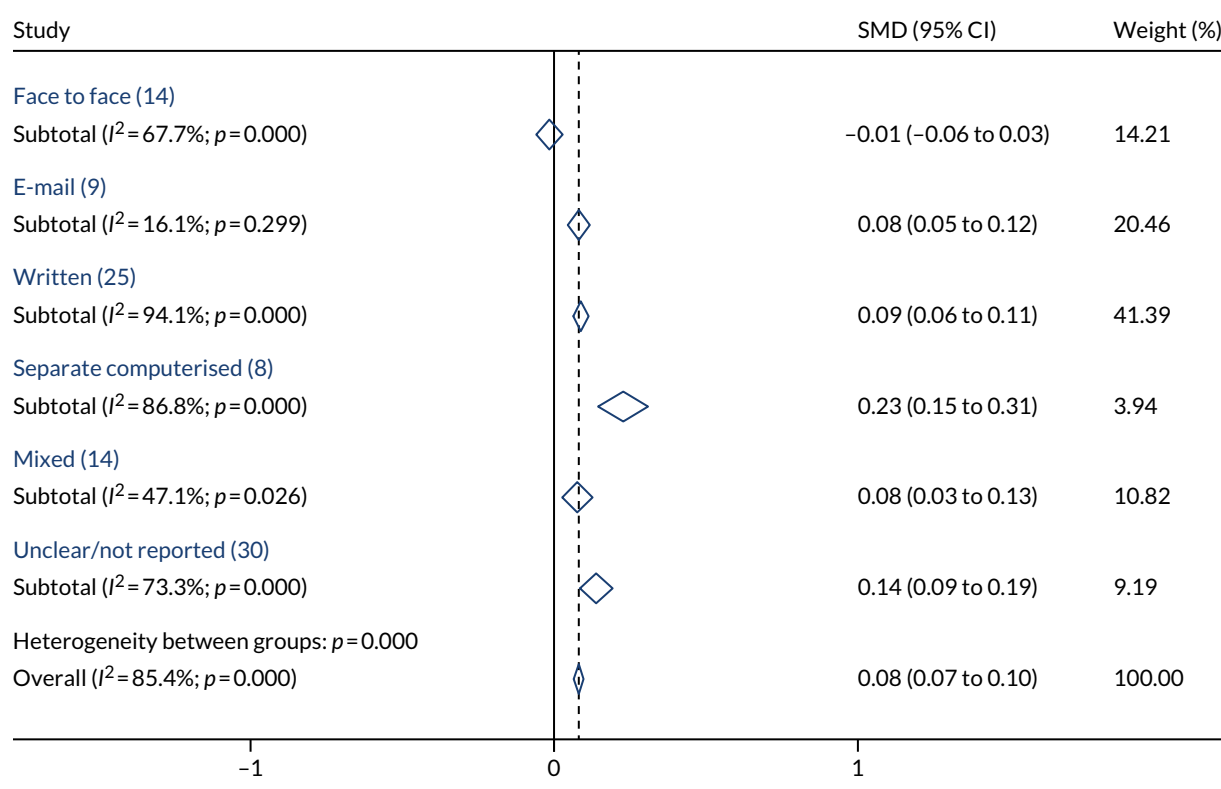


FIGURE 14 Fixed-effects forest plot summarised by format of intervention.

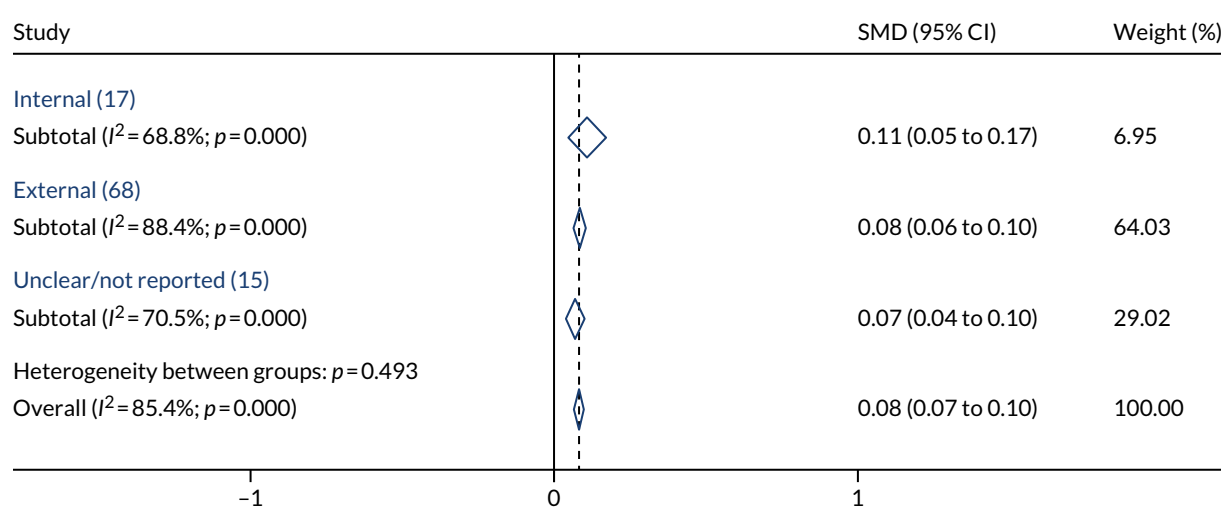


FIGURE 15 Fixed-effects forest plot summarised by person delivering the intervention.

Studies with social norms behaviour change techniques in both arms

There were 17 comparisons in which both arms involved social norms interventions. These are summarised in *Appendix 16, Figure 21*. Most of these involved social comparison in both arms, with feedback on behaviour^{20,70-77} or without feedback¹⁸ ($n = 10$), and typically they were studies that tested some other combination of BCTs in the intervention arm, but had A&F offered as part of usual care in both arms. One study tested a combination of BCTs in the intervention arm, but had credible source in both arms as part of usual care.⁷⁸ None of these studies offered any interesting insights for the review because they were designed to test the effect of other interventions.

Only two studies had a different social norms BCT in each arm, which offered a head-to-head comparison. One study tested the effect of credible source, social comparison and feedback on behaviour and other BCTs against social comparison and feedback, with an estimated SMD of 1.30 (95% CI 0.50 to 2.10),⁷⁹ suggesting that credible source with other BCTs has a high effect compared with social comparison and feedback. Another study found no evidence of a difference between social comparison, credible source and feedback, and credible source alone, with an estimated SMD of 0.29 (95% CI -0.39 to 0.98).⁸⁰

There were four studies that examined the effect of different variants of social comparison ($n = 3$) or credible source ($n = 1$). Wright *et al.*⁸¹ offered formal educational sessions led by a highly regarded surgeon (credible source) to both arms of the trial, and offered one-to-one 'academic detailing' by the highly regarded surgeon to a local opinion leader. The addition of academic detailing was effective compared with the educational sessions alone (SMD 0.32, 95% CI 0.08 to 0.56). Kiefe *et al.*⁸² found that providing social comparison where the average performance for the top 10% of the physicians was reported was more effective than social comparison reporting the mean performance of other physicians (SMD 0.25, 95%CI 0.13 to 1.37). Schneider *et al.*⁸³ similarly tested benchmarked social comparison (best 10% of GPs) with the median performance, and found no evidence of an effect (SMD 0.03, 95% CI -0.05 to 0.56). There was no evidence of a difference in outcome when a similar social comparison intervention was provided to each arm, where the intervention arm was given the information in a work book and the control arm received a graphical computer slide show (SMD -0.04, 95% CI -0.52 to 0.44).⁸⁴

Investigation of behaviour change techniques, settings and contexts, using metaregression

Metaregression: social norms behaviour change techniques and other behaviour change techniques (with feedback on behaviour behaviour change techniques)

For metaregression, we used all of the 100 comparisons that tested the effect of social norms BCTs with two additional comparisons^{79,80} that had different social norms BCTs in each arm. Metaregression allows us to examine the effect of all social norms BCTs plus other common BCTs simultaneously in the same analysis, by using binary covariates for the presence/absence of each BCT in the intervention being tested. When all of the social norms BCTs and other commonly used BCTs are included together in a metaregression (*Table 8*), only credible source stands out as being clearly effective, which suggests that using credible source in an intervention improves compliance with the desired behaviour by an average SMD of 0.29 (95% CI 0.08 to 0.50). Note that there is a large amount of heterogeneity, and that after taking into account the effect of BCTs the residual variation owing to heterogeneity is 85.4%.

Metaregression: social norms and other behaviour change techniques (without feedback on behaviour behaviour change techniques)

Social comparison and feedback on behaviour commonly appear together in the same intervention; therefore, including both of these in the same metaregression would probably cause multicollinearity problems. Repeating the regression reported in *Table 8* but excluding feedback on behaviour gives similar results (*Table 9*), but suggests that social comparison may also have an effect, improving compliance with desired behaviour by an average SMD of 0.12 (95% CI 0.00 to 0.23). The effect of credible source remains fairly consistent (SMD 0.31, 95% CI 0.10 to 0.52). This result is similar to that seen in the forest plots. There is no evidence from this metaregression to suggest that adding other BCTs alongside social comparison or credible source offers any additional improvement once the effect of the social norms BCTs has been taken into account; however, these BCTs were seen only in a small number of trials and the heterogeneity is substantial (residual $I^2 = 85\%$; $\tau^2 = 0.10$), so we must interpret this observation cautiously.

RESULTS

TABLE 8 Results of the metaregression of SMDs for compliance in the desired behaviour using all social norms BCTs plus other commonly used BCTs

Covariate (BCT code)	Effect, SMD (95% CI)	p-value
Social norm BCTs		
Social comparison (6.2)	0.01 (-0.25 to 0.26)	0.96
Credible source (9.1)	0.29 (0.08 to 0.51)	0.01
Social reward (10.4)	0.06 (-0.38 to 0.51)	0.77
Information about others' approval (6.3)	-0.07 (-0.65 to 0.50)	0.80
Social incentive (10.5)	-0.27 (-1.28 to 0.73)	0.59
Other BCTs		
Feedback on behaviour (2.2)	0.14 (-0.15 to 0.42)	0.35
Information about health consequences (5.1)	-0.10 (-0.35 to 0.15)	0.42
Prompts/cues (7.1)	0.05 (-0.17 to 0.27)	0.67
Social support (unspecified) (3.1)	0.07 (-0.14 to 0.28)	0.52
Instruction on how to perform the behaviour (4.1)	-0.05 (-0.27 to 0.17)	0.65

Metaregression: contexts and settings

When including the chosen regression coefficients for factors connected to context and settings, either independently or simultaneously with a constant term, there is no clear evidence that any of these factors are related to treatment effect (*Table 10*). Very little variability has been explained by the inclusion of these covariates (residual $I^2 = 85.6\%$; $\tau^2 = 0.11$). The metaregression conducted was a random-effects metaregression, whereas earlier exploratory forest plots were fixed effect; fixed-effect metaregression is not recommended as a valid method⁸⁵ because it assumes that all heterogeneity can be explained by the regression covariates, and it leads to a high risk of type 1 errors. The difference between random-effect and fixed-effect analyses explains, in part, why subgroups that appeared quite separate on forest plots do not lead to statistically significant covariates in the metaregression. Metaregression is also subject to low power and overfitting, and although we have 102 comparisons included, this may not be sufficient to lead to stable covariate estimates.

TABLE 9 Results of the metaregression of SMDs for compliance in the desired behaviour using social norm BCTs plus other commonly used BCTs, excluding feedback on behaviour (2.2)

Covariate (BCT code)	Effect, SMD (95% CI)	p-value
Social norm BCTs		
Social comparison (6.2)	0.12 (0.00 to 0.23)	0.06
Credible source (9.1)	0.31 (0.10 to 0.52)	0.01
Social reward (10.4)	0.15 (-0.26 to 0.55)	0.48
Information about others' approval (6.3)	-0.17 (-0.71 to 0.37)	0.54
Social incentive (10.5)	-0.06 (-0.95 to 0.83)	0.90
Other BCTs		
Information about health consequences (5.1)	-0.10 (-0.35 to 0.15)	0.43
Prompts/cues (7.1)	0.07 (-0.14 to 0.28)	0.50
Social support (unspecified) (3.1)	0.10 (-0.10 to 0.30)	0.34
Instruction on how to perform the behaviour (4.1)	-0.05 (-0.27 to 0.18)	0.69

TABLE 10 Results of metaregression of SMDs for compliance in the desired behaviour using context, format and settings

Setting/context	Single variable regression, SMD (95% CI)	Multivariable regression, SMD (95% CI)
Health-care worker		
Doctor vs. other	-0.02 (-0.21 to 0.16)	-0.04 (-0.24 to 0.16)
Behaviour		
Prescribing vs. tests	0.10 (-0.12 to 0.33)	0.12 (-0.12 to 0.36)
Management/communication vs. tests	0.08 (-0.18 to 0.34)	0.08 (-0.21 to 0.38)
Other vs. tests	-0.10 (-0.36 to 0.17)	-0.11 (-0.42 to 0.20)
Setting		
Secondary care vs. primary care	0.03 (-0.17 to 0.23)	0.02 (-0.20 to 0.25)
Other vs. primary care	-0.01 (-0.24 to 0.22)	-0.05 (-0.30 to 0.20)
Direction of change required		
Increase vs. decrease	-0.01 (-0.19 to 0.18)	0.00 (-0.23 to 0.23)
Format		
Face to face vs. written	0.04 (-0.23 to 0.32)	0.07 (-0.24 to 0.37)
Computer (including e-mails) vs. written	0.07 (-0.19 to 0.33)	0.07 (-0.21 to 0.34)
Mixed/unclear vs. written	0.09 (-0.12 to 0.30)	0.10 (-0.13 to 0.34)

Investigation of behaviour change techniques, settings and contexts, using network meta-analysis

Network meta-analysis is a suitable method to rank the social norm BCTs from most effective to least effective. For the network meta-analysis, we used all of the 100 comparisons that tested the effect of social norm BCTs with two additional comparisons^{79,80} that had different social norm BCTs in each arm. Some regrouping was carried out to the social norm categories to reduce the number of categories and avoid small groups: all comparisons testing social reward with or without other BCTs were combined, all comparisons testing credible source with or without other BCTs were combined and all comparisons testing social comparison and credible source with or without other BCTs were combined. The number of comparisons available for each of the social norm BCTs is shown in *Table 11*, and the diagram of how they are networked is shown in *Figure 16*.

Network meta-analysis gives similar effect sizes and CIs to those seen in the meta-analysis (see *Figure 3*), but also allows us to rank the social norms interventions from best to worst (*Table 12*). The evidence from 102 tests of social norm BCTs suggests that the most effective interventions contain social comparison and social reward (SMD 0.39, 95% CI 0.15 to 0.64), social comparison and prompts/cues (SMD 0.33, 95% CI 0.22 to 0.44) or credible source (SMD 0.30, 95% CI 0.13 to 0.47). Social comparison on its own (SMD 0.05, 95% CI 0.03 to 0.08) or combined with social support (unspecified) (SMD 0.10, 95% CI 0.04 to 0.16) or other BCTs (SMD 0.04, 95% CI 0.00 to 0.08) all appear to be more effective than control, on average, in improving compliance with the desired behaviour; however, they were associated with a very modest effect size. The use of credible source and social comparison together (SMD 0.16, 95% CI 0.12 to 0.20), and other combinations of two or more social norm BCTs together (SMD 0.07 95% CI 0.03 to 0.12), similarly have a modest effect on behavioural outcomes. There is no evidence to suggest that social reward (SMD 0.03, 95% CI -0.08 to 0.13), social comparison, instruction on how to perform the behaviour and prompts/cues (SMD 0.01, 95% CI -0.10 to 0.11), or social comparison and information about health consequences (SMD -0.14, 95% CI -0.33 to 0.05) offer benefit above control; however, in all cases the CIs were wide and we cannot rule out a modest effect.

RESULTS

TABLE 11 Number of comparisons for each social norm BCT used in network meta-analysis

Intervention group	Type of control group		Total
	0. Control	1. Social comparison	
1. Social comparison	33	N/A	33
2. Credible source	7	0	7
3. Social reward	2	0	2
4. Social comparison and credible source	8	2	10
5. Social comparison and social reward	2	0	2
6. Social comparison and social support (unspecified)	7	0	7
7. Social comparison and prompts/cues	5	0	5
8. Social comparison and information about health consequences	4	0	4
9. Social comparison and instruction on how to perform the behaviour and prompts/cues	5	0	5
10. Social comparison and other BCTs	23	0	23
11. Other multiple social norm BCTs	4	0	4
Total	100	2	102

N/A, not applicable.

Note that numbering follows that used in Figure 16.

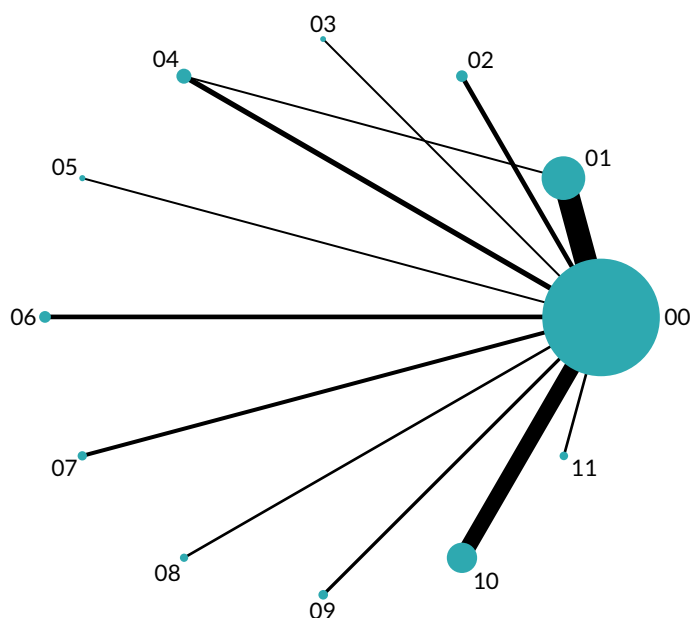


FIGURE 16 Network diagram to showing the available comparisons. 0, any control; 1, social comparison; 2 credible source; 3, social reward; 4, social comparison and credible source; 5, social comparison and social reward; 6, social comparison and social support (unspecified); 7, social comparison and prompts/cues; 8, social comparison and information about health consequences; 9, social comparison and instruction on how to perform the behaviour and prompts/cues; 10, social comparison and other BCTs; 11, other multiple social norm BCTs.

TABLE 12 Intervention effects calculated from network meta-analysis, ordered by effect size

Effect	Effect, SMD (95% CI)	Probability of being the best intervention (%)
Social comparison and social reward vs. control	0.39 (0.15 to 0.64)	59.2
Social comparison and prompts/cues vs. control	0.33 (0.22 to 0.44)	22.2
Credible source vs. control	0.30 (0.13 to 0.47)	18.6
Social comparison and credible source vs. control	0.16 (0.12 to 0.20)	0.0
Social comparison and social support (unspecified) vs. control	0.10 (0.04 to 0.16)	0.0
Other multiple social norm BCTs vs. control	0.07 (0.03 to 0.12)	0.0
Social comparison vs. control	0.05 (0.03 to 0.08)	0.0
Social comparison and other BCTs vs. control	0.04 (0.00 to 0.08)	0.0
Social reward vs. control	0.03 (-0.08 to 0.13)	0.0
Social comparison, instruction ^a and prompts/cues vs. control	0.01 (-0.10 to 0.11)	0.0
Social comparison and information on health consequences vs. control	-0.14 (-0.33 to 0.05)	0.0

a Instruction on how to perform the behaviour.

Sensitivity analyses

Tables 13 and 14 show that our main conclusion is robust: there is little change in our average SMD when we impute alternative values for the ICC when required, exclude trials in which the standard deviation was imputed, remove trials reporting mean per cent compliance in which the mean compliance was close to 0% or 100%, or include trials at only low risk of bias on each domain. Note that we decided before analysis not to do a sensitivity analysis on 'risk of bias due to blinding', as we believe blinding to be difficult/impossible in social norm interventions and, as expected, it was rarely seen; however, we must bear in mind that our review is at risk of bias owing to this lack of blinding.

TABLE 13 Sensitivity analysis for overall result, fixed effects

Analysis	Effect, SMD (95% CI)	Number of comparisons
Full data set	0.08 (0.07 to 0.10)	100
Using imputed ICC = 0.2 instead of 0.1	0.08 (0.06 to 0.10)	100
Using imputed ICC = 0.05 instead of 0.1	0.09 (0.07 to 0.10)	100
Removing trials with imputed SDs	0.09 (0.07 to 0.10)	94
Removing trials reporting mean per cent compliance close to 0% or 100% ^a	0.07 (0.05 to 0.08)	77
Keeping only trials at low risk of bias owing to allocation concealment	0.11 (0.09 to 0.13)	72
Keeping only trials at low risk of bias owing to sequence generation	0.08 (0.06 to 0.10)	74
Keeping only trials at low risk of bias owing to selective outcome reporting	0.09 (0.07 to 0.11)	41
Keeping only trials at low risk of bias owing to attrition	0.10 (0.08 to 0.12)	57
Keeping only trials at low risk of bias owing to other biases	0.09 (0.07 to 0.11)	59
Removing trials in which 'feedback on desired behaviour' was not part of the tested intervention	0.08 (0.07 to 0.10)	88

SD, standard deviation.

a Trials using mean per cent compliance and reporting mean per cent compliance < 20% or > 80%.

TABLE 14 Sensitivity analysis for overall result, random effects

Analysis	Effect, SMD (95% CI)	Number of comparisons
Full data set	0.16 (0.11 to 0.22)	100
Using imputed ICC = 0.2 instead of 0.1	0.16 (0.10 to 0.21)	100
Using imputed ICC = 0.05 instead of 0.1	0.16 (0.11 to 0.21)	100
Removing trials with imputed SDs	0.18 (0.12 to 0.23)	94
Removing trials reporting mean per cent compliance close to 0% or 100% ^a	0.12 (0.07 to 0.16)	77
Keeping only trials at low risk of bias owing to allocation concealment	0.18 (0.12 to 0.25)	72
Keeping only trials at low risk of bias owing to sequence generation	0.17 (0.10 to 0.23)	74
Keeping only trials at low risk of bias owing to selective outcome reporting	0.22 (0.13 to 0.31)	41
Keeping only trials at low risk of bias owing to attrition	0.18 (0.13 to 0.24)	57
Keeping only trials at low risk of bias owing to other biases	0.13 (0.09 to 0.18)	59
Removing trials in which 'feedback on desired behaviour' was not part of the tested intervention	0.17 (0.12 to 0.23)	88

SD, standard deviation.

^a Trials using mean per cent compliance and reporting mean per cent compliance < 20% or > 80%.

In an additional unplanned sensitivity analysis, we excluded comparisons that did not include feedback on behaviour alongside the social norm BCT being tested. Conclusions would be similar whether or not comparisons that included feedback on behaviour were included.

Publication bias

There is some evidence of funnel plot asymmetry, with several SMDs lying on the right-hand side outside the predicted funnels (*Figure 17*). This means that the review may be missing some unpublished negative trials, or may include more positive trials than justified owing to selective outcome reporting. We should review the results cautiously in the light of the risk of outcome reporting bias, especially when we look at the magnitude of the extreme positive trials in relation to the overall treatment effect.

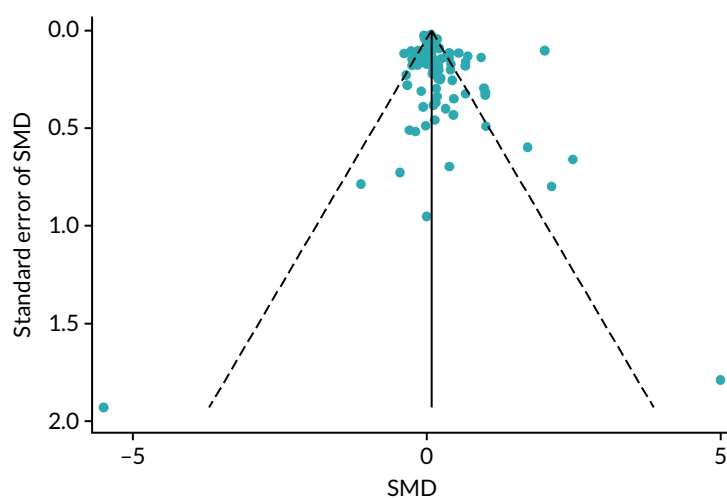


FIGURE 17 Funnel plot: the intervention effect estimates from individual studies against the standard error.

Effects of interventions: patient health outcomes (secondary)

Figure 18 shows the SMD among patient outcomes (14 comparisons), grouped by type of BCT comparison, in a fixed-effects meta-analysis. Only a subset of comparison types is represented compared with the results for primary (health worker behaviour) outcomes, given that not all studies reported a patient outcome. As for health worker behaviour outcomes, heterogeneity is high with an overall $I^2 = 91.5\%$. Combined data from these 14 comparisons suggest that interventions with a social norm component were associated with an improvement in patient outcomes of 0.17 SMD (95% CI 0.14 to 0.20), on average. However, this is strongly influenced by those studies testing social comparison, in particular Bentz *et al.*⁸⁷ (weight 46%) with an estimated SMD of 0.36 (95% CI 0.32 to 0.40) and Beck *et al.*⁸⁶ (weight 31%) with an estimated SMD of 0.00 (95% CI -0.05 to 0.05). Estimates consistent with a null effect were found for all studies testing social comparison combined with social support (unspecified), prompts/cues, instruction on the behaviour plus prompts/cues or other BCTs. A larger positive effect of 0.86 SMD (95% CI 0.29 to 1.44) is found for the test of a credible source intervention, but this is from one small study (weight 0.3%).⁶⁷ These results for patient outcomes should be interpreted cautiously owing to a large amount of heterogeneity among studies, and to the small number of studies in some groups.

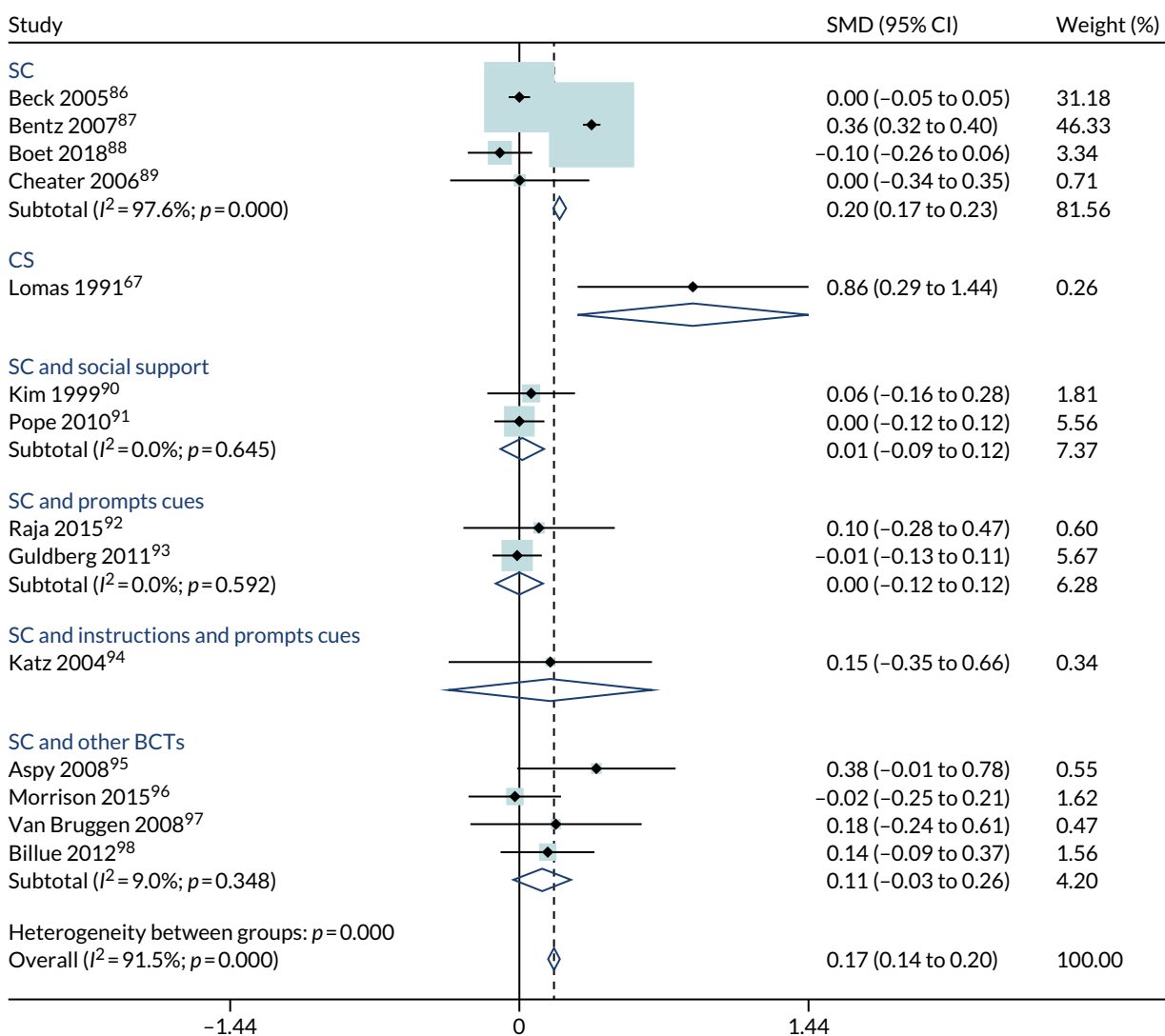


FIGURE 18 Patient outcomes: fixed effects summarised by type of comparison (15 comparisons). CS, credible source; SC, social comparison; SR, social reward.

RESULTS

Both the patient outcomes and the target behaviours varied across the studies. To illustrate the findings, we offer some examples of studies included in the review. Ivers *et al.*¹⁸ targeted GP behaviour in arranging testing and prescribing for people with diabetes, and the patient outcome was mean systolic blood pressure. Aspy *et al.*⁹⁵ targeted primary care physicians, encouraging them to offer a mammogram to appropriate women patients, and the patient outcome was the proportion of eligible patients who had a mammogram. Lomas *et al.*⁶⁷ targeted secondary care doctors, encouraging them to reduce the offer of caesareans to women who had previously had caesareans, and the patient outcome was the proportion of vaginal births. Billue *et al.*⁹⁸ targeted GP behaviour in intensifying medication for a range of health conditions, and the patient outcome was the proportion of patients with controlled diabetes. In studies in which there was more than one health outcome, we chose the study's primary outcome, and if that was not clear, we chose the first outcome that was reported.

Figure 19 shows the SMD among patient outcomes (14 comparisons), grouped by type of BCT comparison, using a random-effects meta-analysis. As for the fixed-effects analysis, only a subset of comparison types is represented compared with the results for primary outcomes, given that not all studies reported a patient outcome.

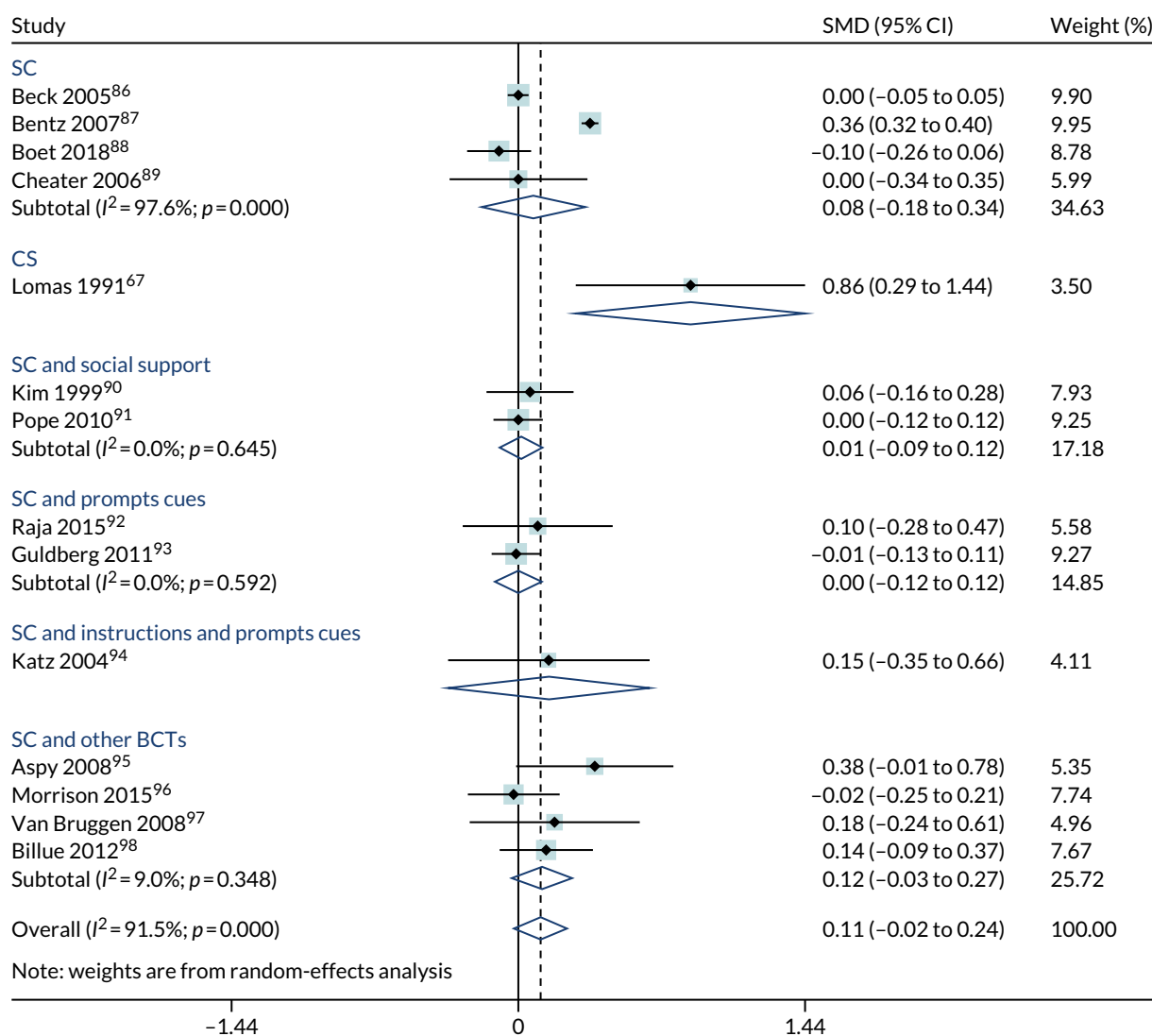


FIGURE 19 Patient outcomes: random effects grouped by type of comparison (14 comparisons). CS, credible source; SC, social comparison; instructions, instruction on how to perform the behaviour; social support, social support (unspecified).

The overall estimate of effect is less precise and slightly attenuated compared with the corresponding fixed-effects analysis, with an overall estimated SMD of 0.11 (95% CI -0.02 to 0.24). The results for 12 out of the 14 comparisons are consistent with no effect, exceptions were Lomas *et al.*'s⁶⁷ (weight 3.5%) test of a credible source intervention, with an estimated SMD of 0.86 (95% CI 0.29 to 1.44) and Bentz *et al.*'s⁸⁷ (weight 10%) test of a pure social comparison intervention, with an estimated SMD of 0.36 (95% CI 0.32 to 0.40). However, the overall estimated group SMD for pure social comparison interventions is 0.08 (95% CI -0.18 to 0.34), consistent with no effect. As for the corresponding fixed-effects analysis, the results should be interpreted cautiously owing to a high amount of heterogeneity and relatively few comparisons in each group.

6.6 PAPER 6

Tang MY, Rhodes S, Powell R, McGowan L, Howarth E, Brown B, Cotterill S. How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis. *Implementation Science*. 2021;16(1):8.


<https://doi.org/10.1186/s13012-020-01072-1>

SYSTEMATIC REVIEW

Open Access



How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis

Mei Yee Tang^{1,2*} , Sarah Rhodes¹, Rachael Powell³, Laura McGowan³, Elizabeth Howarth¹, Benjamin Brown^{4,5} and Sarah Cotterill¹

Abstract

Background: Healthcare workers perform clinical behaviours which impact on patient diagnoses, care, treatment and recovery. Some methods of supporting healthcare workers in changing their behaviour make use of social norms by exposing healthcare workers to the beliefs, values, attitudes or behaviours of a reference group or person. This review aimed to evaluate evidence on (i) the effect of social norms interventions on healthcare worker clinical behaviour change and (ii) the contexts, modes of delivery and behaviour change techniques (BCTs) associated with effectiveness.

Methods: Systematic review and meta-analysis of randomised controlled trials. Searches were undertaken in seven databases. The primary outcome was compliance with a desired healthcare worker clinical behaviour and the secondary outcome was patient health outcomes. Outcomes were converted into standardised mean differences (SMDs). We performed meta-analyses and presented forest plots, stratified by five social norms BCTs (*social comparison, credible source, social reward, social incentive and information about others' approval*). Sources of variation in social norms BCTs, context and mode of delivery were explored using forest plots, meta-regression and network meta-analysis.

(Continued on next page)

* Correspondence: meiyee.tang@newcastle.ac.uk

¹Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester M13 9PL, UK

²National Institute of Health Research Behavioural Science Policy Research Unit, Population Health Sciences, Baddiley-Clark Building, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne NE2 4AX, UK

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

Results: Combined data from 116 trials suggested that social norms interventions were associated with an improvement in healthcare worker clinical behaviour outcomes of 0.08 SMDs (95%CI 0.07 to 0.10) ($n = 100$ comparisons), and an improvement in patient health outcomes of 0.17 SMDs (95%CI 0.14 to 0.20) ($n = 14$), on average. Heterogeneity was high, with an overall I^2 of 85.4% (healthcare worker clinical behaviour) and 91.5% (patient health outcomes). *Credible source* was more effective on average, compared to control conditions (SMD 0.30, 95%CI 0.13 to 0.47, $n = 7$). *Social comparison* also appeared effective, both on its own (SMD 0.05, 95%CI 0.03 to 0.08, $n = 33$) and with other BCTs, and seemed particularly effective when combined with *prompts/cues* (0.33, 95%CI 0.22 to 0.44, $n = 5$).

Conclusions: Social norms interventions appeared to be an effective method of changing the clinical behaviour of healthcare workers and have a positive effect on patient health outcomes in a variety of health service contexts. Although the overall result is modest and variable, there is the potential for social norms interventions to be applied at large scale.

Trial registration: PROSPERO [CRD42016045718](https://doi.org/10.1186/1745-6215-42016045718).

Keywords: Systematic review, Meta-analysis, Health professional behaviour, Social norm, Social comparison, Information about others' approval, Credible source, Social reward, Social incentive, Audit and feedback

Contributions to the literature

- This is the first systematic review and meta-analysis on the use of social norms interventions to change the clinical behaviour of healthcare workers, and the results suggest that, on average, these interventions are effective.
- Social norms interventions may be effective across a range of health service contexts and modes of delivery, but the effects are variable.
- These findings contribute to a recognised gap in the literature, by highlighting which social norms interventions may be most effective: this can inform the design of future interventions aimed at improving health professional practice.

Background

Healthcare workers routinely perform behaviours in clinical settings which impact all aspects of patient care including diagnoses, treatment and recovery. There are best-practice guidelines for many of these clinical behaviours. For example, regular blood glucose testing for diabetic patients. Healthcare workers face many challenges in following evidence-based professional practice such as lack of time, competing demands and requests from patients. Although there are no reliable published estimates of how well healthcare workers follow best clinical practices, 1 in 20 hospital admissions is caused by adverse drug events [1], and approximately half of these globally are believed to be due to lapses in best practice in terms of prescribing or monitoring behaviours by clinicians [2].

Social influences are important in clinical practice: prescribers of antibiotics have reported that pressure

from patients and other prescribers in their networks influence their prescribing behaviours [3]. Social norms can be broadly considered as the perceived implicit or explicit behavioural rules that one uses to determine the appropriate and/or typical expectations, beliefs, attitudes and behaviours of a social reference person or group [4]. We have defined a social norms intervention as one which seeks to change the clinical behaviour of a target healthcare worker by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person. The target healthcare worker is the person at whom a social norms intervention is aimed, with a view to changing their clinical behaviour. The reference person or group describes a person or group whose values, beliefs or behaviours are exposed to the target. Social norms interventions sometimes report a peer benchmark, such as the top 10% of the reference group or the average performance: the downside of the average approach is that the above-average performers will receive feedback suggesting that they are already performing better than their peers, and this may lead them to reduce their effort [5].

Behaviour change interventions based on social norms may help overcome barriers to healthcare workers implementing recommended practice through: persuasion, encouraging collaboration to achieve change, observing good practice from elsewhere and support from management [6]. There are various explanations of the processes through which social norms impact on behaviour according to social and health psychology theories. Social comparison theory [7] proposes that individuals draw on social comparisons to evaluate one's abilities and perform behaviours which will bring one's abilities in line with those of others in the group. According to the social identity perspective [8], people make

evaluations about their own group ('in group norms') against other groups ('out group norms'). They are motivated to preserve their social identity (as part of their 'in group') by behaving in similar ways to the group's normative behaviour. 'Subjective norm' is a construct within the Theory of Planned Behaviour [9], which describes an individual's perception of whether valued others think they should perform a behaviour, combined with a motivation to comply with others' beliefs.

A social norms intervention with a descriptive norms [10] message provides the target with information about the behaviour of others in the reference group (such as providing a nurse with information about the behaviours of nurses regarding wound dressing). An injunctive norms message provides the target worker with information about the values, beliefs or attitudes of the reference group towards a particular behaviour, conveying social approval or disapproval (e.g. saying that colleagues disapprove of ordering unnecessary tests). This includes approval, praise, commendation, applause or thanks.

Audit and feedback (A&F) is a quality improvement technique used by health services, where data is collected on healthcare worker performance and then a summary is reported back to the individual [11]. Social norms interventions are sometimes included as one component of A&F, usually by providing descriptive norms of others' behaviour [12, 13]. A&F has already been shown to be effective in changing healthcare worker behaviour, but with large variation in outcomes depending on the context and the intervention design [14]. There is a need to understand the ingredients for successful A&F [11, 15], and the effects or mechanisms of the 'social influence' constituents of A&F have been identified as topics for further research [11]. Our review contributes to this important research agenda by systematically examining the evidence for using social norms interventions with healthcare workers.

Identification of the individual components within social norms interventions can aid understanding of the precise aspects that influence behaviour. The Behaviour Change Techniques Taxonomy v1 (BCTTv1) [16] is a framework for classifying BCTs, which are the 'active ingredients' of behaviour change interventions. The taxonomy defines 93 distinct BCTs, grouped into categories. There is no explicit category that relates to social norms. For this review, five BCTs were considered to involve social norms: '6.2. *Social Comparison*', '6.3. *Information about Others' Approval*', '9.1. *Credible Source*', '10.4. *Social Reward*', and '10.5. *Social Incentive*'. The numbers follow the BCTTv1 labelling and definitions are listed in Table 1.

The aim was to conduct a systematic review to assess the impact on healthcare workers' compliance with professional practice recommendations of interventions

delivering social norms BCTs, compared to controls. Two research questions were addressed:

1. What is the effect of interventions containing social norms BCTs on (a) the clinical behaviour of healthcare workers, and (b) resulting patient health outcomes?
2. Which contexts, modes of delivery and behaviour change techniques are associated with the effectiveness of social norms interventions on healthcare worker clinical behaviour change?

Methods

The study design was a systematic review with meta-analysis [18], meta-regression [19] and network meta-analysis [20]. This paper follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [21]. Six members of the public attended workshops to discuss the relevance of the review to patients and carers, study design and dissemination. The group felt that patients can potentially have a role in changing healthcare worker behaviour, for example by reminding healthcare workers to wash their hands; or telling the General Practitioner (GP) they do not want antibiotics for a cold, although they were cynical about whether doctors would listen. In response, we changed our data collection to record whether any studies considered patients' role in social norms interventions. Their advice on how to interpret our results to a broad audience will influence our future dissemination plans. An independent study steering committee, including a member of the public, provided encouragement and counsel throughout the project.

Protocol and registration

The study was registered on PROSPERO (CRD42016045718) and a protocol is available [17].

Searches

A search strategy was developed, following an iterative process of scoping searches. In July 2018, searches were undertaken in MEDLINE, PsycINFO, EMBASE, CINAHL, BNI, Cochrane CENTRAL and Web of Science (see Appendix 1). Backward and forward citation searching was not conducted, as per the protocol, due to time constraints.

Study inclusion criteria

Studies were included if they met the criteria in Table 2.

Screening

Covidence was used to facilitate screening and data extraction [22]. One reviewer screened all titles and abstracts against the inclusion criteria; a second reviewer

Table 1 Definitions of social norms (SN) BCTs and other (non-SN) BCTs that feature prominently in SOCIAL review

SN/non-SN BCT	Name and definition from BCT taxonomy (reproduced from the BCT taxonomy [16])	SOCIAL review name and definition (reproduced from the SOCIAL protocol [17])
Social norm BCT	<p>6.2. Social comparison Draw attention to others' performance to allow comparison with the person's own performance. Note: being in a group setting does not necessarily mean that social comparison is actually taking place. Show the doctor the proportion of patients who were prescribed antibiotics for a common cold by other doctors and compare with their own data.</p> <p>6.3. Information about others' approval Provide information about what other people think about the behaviour. The information clarifies whether others will like, approve or disapprove of what the person is doing or will do. Tell the staff at the hospital ward that staff at all other wards approve of washing their hands according to the guidelines.</p>	Coded as per original definition, unchanged.
Social norm BCT	<p>9.1. Credible source Present verbal or visual communication from a credible source in favour of or against the behaviour. Note: code this BCT if source generally agreed on as credible, e.g. health professionals, celebrities or words used to indicate expertise or leader in field and if the communication has the aim of persuading. Present a speech given by a high-status professional to emphasise the importance of not exposing patients to Unnecessary radiation by ordering X-rays for back pain.</p>	Coded as per original decision, unchanged.
Social norm BCT	<p>10.4. Social reward Arrange verbal or non-verbal reward if and only if there has been effort and/or progress in performing the behaviour (includes 'positive reinforcement'). Congratulate the person for each day they eat a reduced fat diet.</p>	Coded as per original decision, unchanged.
Social norm BCT	<p>10.5 Social incentive Inform that a verbal or non-verbal reward will be delivered if and only if there has been effort and/or progress in performing the behaviour (includes 'positive reinforcement'). Inform that they will be congratulated for each day that they eat a reduced fat diet.</p>	<p>Changed: Inform that praise, commendation, applause or thanks will be delivered if and only if there has been effort and/or progress in performing the behaviour (includes 'positive reinforcement'). New example, relevant to healthcare worker context: arrange for a family doctor to be sent a thank you note for each week that they reduce their level of antibiotic prescribing. Reason for change: the definition of social reward as 'verbal or non-verbal reward' is insufficient to distinguish a 'social' reward from other types of reward. Further, in the present study, we are interested in only those social rewards that rely on social norms. Praise, commendation, applause or thanks are all injunctive norms messages, providing the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval.</p>
Other BCT (not social norm)	<p>7.1. Prompts and cues Introduce or define environmental or social stimulus with the purpose of prompting or cueing the behaviour. The prompt or cue would normally occur at the time or place of</p>	<p>Changed: Inform that praise, commendation, applause or thanks will be delivered if and only if there has been effort and/or progress in performing the behaviour (includes 'positive reinforcement'). New example, relevant to healthcare worker context: Promise a family doctor in advance that they will be sent a thank you note for each week that they reduce their level of antibiotic prescribing. Reason for change The definition of social reward as 'verbal or non-verbal reward' is insufficient to distinguish a 'social' reward from other types of reward. Further, in the present study, we are interested in only those social rewards that rely on social norms. Praise, commendation, applause or thanks are all injunctive norms messages, providing the target with information about the values, beliefs or attitudes of the reference group, conveying social approval or disapproval.</p> <p>Coded as per original definition, unchanged.</p>

Table 1 Definitions of social norms (SN) BCTs and other (non-SN) BCTs that feature prominently in SOCIAL review (Continued)

SN/non-SN BCT	Name and definition from BCT taxonomy (16)	SOCIAL review name and definition (reproduced from the SOCIAL protocol (17))
Other BCT (not social norm)	<p>performance Note: when a stimulus is linked to a specific action in an if-then plan including one or more of frequency, duration or intensity also code 1.4, <i>Action planning</i>.</p> <p>Put a sticker on the bathroom mirror to remind people to brush their teeth</p> <p>3.1. <i>Social support (Unspecified)</i></p> <p>Advise on, arrange or provide social support (e.g. from friends, relatives, colleagues, buddies or staff) or non-contingent praise or reward for performance of the behaviour. It includes encouragement and counselling, but only when it is directed at the behaviour.</p> <p>Note: attending a group class and/or mention of 'follow-up' does not necessarily apply this BCT; support must be explicitly mentioned; if practical, code 3.2, <i>Social support (practical)</i>; if emotional, code 3.3, <i>Social support (emotional)</i> (includes 'Motivational interviewing' and 'Cognitive Behavioural Therapy').</p> <p>Advise the person to call a 'buddy' when they experience an urge to smoke.</p> <p>Arrange for a housemate to encourage continuation with the behaviour change programme.</p> <p>Give information about a self-help group that offers support for the behaviour.</p>	Coded as per original definition, unchanged.
Other BCT (not social norm)	<p>4.1. <i>Instructions on how to perform the behaviour</i></p> <p>Advise or agree on how to perform the behaviour (includes 'Skills training'). Note: when the person attends classes such as exercise or cookery, code 4.1, <i>Instruction on how to perform the behaviour</i>, 8.1, <i>Behavioural practice/rehearsal and 6.1, Demonstration of the behaviour</i>.</p> <p>Advise the person how to put a condom on a model of a penis correctly</p>	Coded as per original definition, unchanged.
Other BCT (not social norm)	<p>5.1. <i>Information on Health Consequences</i></p> <p>Provide information (e.g. written, verbal, visual) about health consequences of performing the behaviour. Note: consequences can be for any target, not just the recipient(s) of the intervention; emphasising importance of consequences is not sufficient; if information about emotional consequences, code 5.6, <i>Information about emotional consequences</i>; if about social, environmental or unspecified consequences code 5.3, <i>Information about social and environmental consequences</i>.</p>	Coded as per original definition, unchanged.

Table 2 Inclusion criteria

PICOS criterion	Description
Population	Healthcare workers, including managers and those in training.
Intervention	A social norms intervention in a (non-simulated) healthcare setting that seeks to change the clinical behaviour of target population by exposing them to the values, beliefs, attitudes, or behaviours of a reference group or person.
Comparison/control	No restrictions on the comparators.
Outcomes	Primary outcome of interest was compliance with the desired clinical behaviour. Secondary outcomes were patient health-related outcomes.
Study design	Randomised controlled trials published in peer-reviewed journals, in English Language. Grey literature was not eligible for inclusion.

screened a 20% random sample to assess reliability. Studies included to the full-text stage were independently screened by two researchers. Any disagreements were resolved through discussion, moderation of a third researcher or team review.

Data extraction

Data from included studies were extracted using a tailored data extraction form (Appendix 2) [23]. Information relating to the population and setting, methods, participant characteristics, intervention characteristics (delivery and BCT content), comparators, outcomes and results were extracted.

For the primary outcome (healthcare worker clinical behaviour), we extracted all available summary data on compliance of the healthcare worker with the desired behaviour at the time point closest to 6 months post-randomisation. Where multiple measures of compliance were reported we followed this list of priorities: (a) reported in sufficient detail to calculate standardised mean difference, (b) observed rather than self-report, (c) appropriate adjustment for clustering, (d) continuous measure, (e) final score rather than change from baseline, (f) described as primary outcome, (g) used to calculate sample size and (h) reported first. A similar approach was followed for patient health outcomes.

All identified BCTs (including both social norms and non-social norms) in all control and intervention arms of included studies were independently coded by two trained researchers using the BCTTv1 [16] and recorded on a BCT extraction form (Appendix 3). The intervention descriptions from all relevant papers (including protocols, process evaluations or additional sources cited in the included studies) were coded to capture the BCTs as closely as possible. Inter-rater reliability for each of the BCTs that were present at least once across all arms was assessed using the prevalence and bias-adjusted kappa (PABAK) statistic (see Appendix 4), which adjusts for both the prevalence and occurrence of BCTs [24]. In circumstances where prevalence is low, the widely used chance-corrected kappa statistic is likely to

underestimate reliability as it is highly dependent on prevalence [25].

Study quality assessment

Risk of bias was independently assessed by two researchers using the Cochrane Collaboration risk of bias tool. The percentages of high/low/unclear judgements for each criterion across included studies were calculated.

Data analysis/synthesis

Any observed measure of healthcare worker behaviour was converted into a standardised mean difference (SMD, Cohen's *D*) comparing intervention and control groups [26]. Odds ratios were converted to SMDs [27]. Where necessary, the sign of the SMD was changed to ensure that a positive SMD represented an improvement in compliance with the desired behaviour.

Where data were from appropriately analysed cluster randomised trials or stepped wedge trials the reported adjusted standard errors were used. Where adjusted standard errors were not reported, we inflated them ourselves to account for clustering [28].

Where data were missing, we searched for companion papers. Missing standard deviations were estimated using any available information (e.g. *p* values, confidence intervals, range, interquartile range) or by searching for trials with similar outcome measures. For cluster randomised trials, we estimated the intraclass correlation coefficient (ICC) where necessary by taking the average of results from similar studies.

Where studies, including factorial trials, assessed more than one intervention, data were extracted for any comparisons that were relevant to the review, avoiding double-counting by dividing the number of participants in the control arm evenly between comparisons. Where there was more than one control arm, the comparison that was the purest test of a social norms intervention was utilised. Where a study was an appropriately analysed factorial trial the covariate and standard error that best estimated the effect of a social norms intervention was extracted.

All studies that reported a primary or secondary outcome measure that could be converted into an SMD were included in meta-analyses. The approach to utilising the five social norms BCTs in the analysis was to subtract the control arm BCTs from those in the intervention arm, to identify those BCTs that were the active ingredients being tested in the trial. The BCT *feedback on behaviour* was present alongside a social norm BCT in 88 of 100 comparisons and so we combined *feedback on behaviour* with the social norm BCT with which it appeared for the purpose of primary meta-analyses.

Fixed effects meta-analysis [29] and forest plots, stratified by BCT were used to assess the effect of social norms on the clinical behaviour of healthcare workers and patient health outcomes. Sources of variation in the type of social norm, context and mode of delivery were explored using both exploratory subgroup analysis and meta-regression [30]. Network meta-analysis [20] was used to (a) utilise all available data and therefore maximise power by including trials that compared two or more different types of social norms (in addition to those that compared a social norm

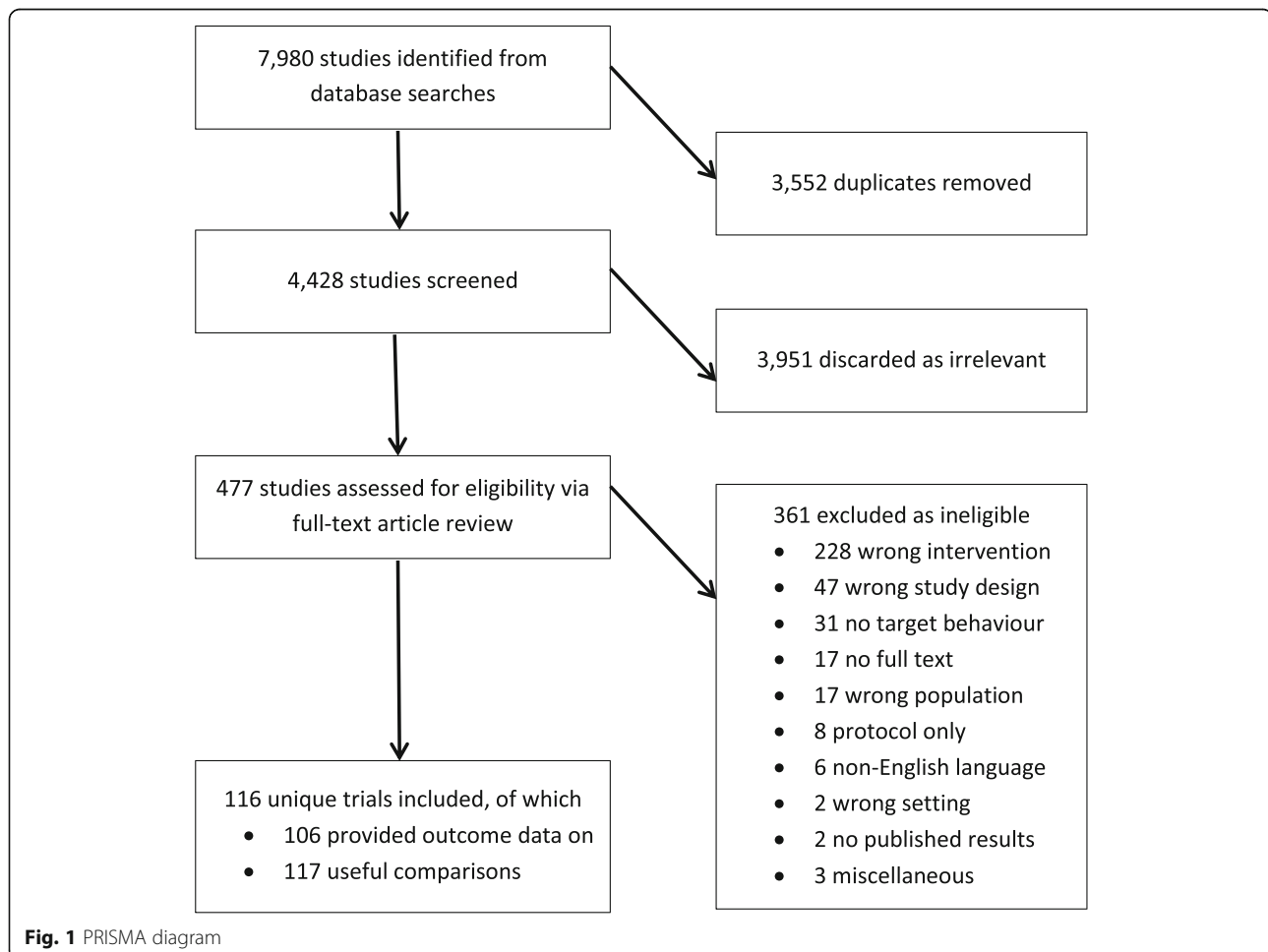
intervention to a control) and (b) rank the different types of social norms intervention in order of effectiveness. A fixed effects approach to meta-analysis was adopted to yield a summary of the evidence in these trials (i.e. the average effect), rather than an estimate of a common underlying treatment effect. Random effects analyses are also reported.

Pre-planned sensitivity analyses assessed the robustness of the conclusions by excluding studies: at high risk of bias on key domains (allocation concealment, sequence generation, selective outcome reporting, attrition, other biases); with ‘mean percentage’ < 20% or > 80% (due to expected skewed distribution) with imputed standard deviations; using estimated ICCs; with and without feedback on desired behaviour.

Results

Study characteristics

There were 4428 citations screened at the title and abstract stage; 477 full-text papers were screened, of which 116 unique trials met the inclusion criteria. Ten of these trials did not report usable outcome



data; therefore, a total of 106 trials contributed findings to the review (Fig. 1, Appendix 5). Some studies had more than one trial arm, resulting in 117 included comparisons. The trial and intervention characteristics are summarised in Table 3, and characteristics of each individual comparison are provided in Appendix 6. There were 100 comparisons suitable for meta-analysis. These included studies testing *social comparison* ($n = 79$) *credible source* ($n = 7$) and *social reward* ($n = 2$) against control. Other studies tested more than one social norm together: *social comparison* and *credible source* ($n = 6$), *social comparison* and *social reward* ($n = 2$), multiple social norms (more than two) together ($n = 4$). Over half of

the included trials were conducted in North America; most studies were set in primary care and hospitals, targeting doctors. A broad range of behaviours were targeted including prescribing, management of conditions and test ordering. Two thirds of the trials were cluster RCTs. The interventions were delivered in a variety of formats; a third was delivered on one occasion and the rest on multiple occasions. Most were delivered by someone outside of the target organisation, often an investigator, and three quarters aimed to increase, rather than decrease the behaviour. Some intervention characteristics were poorly reported; format and frequency of delivery were missing in a third of studies (Table 3).

Table 3 Characteristics of included studies

Study characteristic (n = 106)	No.	%	Study characteristic (n = 106)	No.	%	Intervention characteristic (n = 117)	No.	%
Country			Type of trial			Source		
Australia	8	7.5	Cluster RCT	69	65.1	Peer	6	5.1
Canada	15	14.2	Factorial	4	3.8	Investigators	83	70.9
Denmark	4	3.8	Randomised controlled trial	28	26.4	Supervisor or senior colleague	2	1.7
UK	13	12.3	Stepped wedge	4	3.8	Patient	1	0.9
Netherlands	6	5.7	Matched pairs, cluster RCT	1	0.9	<i>Credible source</i>	15	12.8
USA	45	42.5	Low baseline performance^a			Other	1	0.9
Other/multiple	15	14.2	No	103	97.2	Not reported	9	7.7
Setting			Yes	2	1.9	Internal/external delivery^b		
Primary (GP/GP practice nurses)	57	53.8	Unclear	1	0.9	Internal	17	14.5
Hospital (inpatient and outpatient)	31	29.3				External	81	69.2
Community	4	3.8				Unclear/not reported	19	16.2
Care/nursing home	4	3.8				Reference group		
Mixed	7	6.6				Peer	97	82.9
Other	3	2.8	Intervention characteristic (n = 117)	No.	%	Professional body	1	0.9
Type of HCP			Format			Senior person	9	7.7
Doctor (primary care)	45	42.5	Face-to-face meeting	16	13.7	Patient(s)	1	0.9
Doctor (secondary)	19	17.9	Email	10	8.5	Multiple	4	3.4
Other (nurse/dentist/AHP/pharmacist)	7	6.6	Written (paper)	29	24.8	Unclear/not reported	5	4.3
Mixture/whole team	35	33.0	Separate computerised	10	8.5	Direction of change		
Target behaviour			Mixed	18	15.4	Increase	85	72.6
Prescribing (incl. vaccinations)	40	37.7	Unclear/not reported	34	29.1	Decrease	30	25.6
Handwashing/hygiene	4	3.8	Frequency			Maintenance	0	0.0
Tests/assessments	21	19.8	Only once	35	29.9	Unclear	2	1.7
Referrals	3	2.8	Twice	10	8.5	Comparator		
Management communications	25	23.6	More than twice	45	38.5	Alternative intervention	15	12.8
Other	2	1.9	Unclear/not reported	27	23.1	Usual practice	59	50.4
Multiple	11	10.4				Attention or waitlist control	18	15.4
						Concomitant intervention ^c	25	21.4

^aDoes the inclusion criteria target participants based on low target performance?

^bThe person delivering the intervention internal or external to the target person's organisation?

^cIntervention that appears in both arms

Effects of interventions

Overall effects on clinical behaviours and patient outcomes

Combined data from fixed effects meta-analysis suggested that social norms interventions were associated with an improvement in healthcare worker clinical behaviour of 0.08 SMDs (95%CI 0.07 to 0.10, $n = 100$ comparisons), and an improvement in patient health outcomes of 0.17 SMD (95%CI 0.14 to 0.20), on average. There was a large amount of heterogeneity with an overall I^2 value of 85.4% (primary) and 91.5% (secondary) suggesting that some studies reported substantially higher or lower effects than the average. However, I^2 is related to precision and rapidly approaches 100% when the number of studies is high [31]. Similar conclusions were drawn from random effects meta-analysis an overall improvement in healthcare worker clinical behaviour of 0.16 SMD (95%CI 0.11 to 0.21, $I^2 = 85.4%$, $\tau^2 = 0.043$). Note that the random effects analysis was associated with a larger effect size and wider confidence interval because more weight is given to smaller trials. These results remained robust after all of our pre-planned sensitivity analyses (Appendix 7).

Social norms behaviour change techniques

Meta-analysis, stratified by social norms BCTs indicated that two of the social norms BCTs had a positive effect on healthcare worker clinical behaviour (Fig. 2): *credible source* (with or without other BCTs) (SMD 0.30, 95%CI 0.13 to 0.47, $n = 7$) and *social comparison* (with or without other BCTs) (SMD 0.06, 95%CI 0.04 to 0.08, $n = 77$). *Social reward* may not be effective (SMD 0.03, 95%CI - 0.08 to 0.13, $n = 2$), based on a small sample. We did not find sufficient evidence to examine the effect of the other two social norm BCTs (*information about others' approval* and *social incentive*). Multiple social norms delivered together were also effective on average (SMD 0.13, 95%CI 0.10 to 0.16). When we looked at the most common combinations of social norms BCTs alongside other BCTs, three types of social norms intervention were most effective, on average, compared to control (Table 4): *credible source* (0.30, 95%CI 0.13 to 0.47); *social comparison* combined with *social reward* (0.39, 95%CI 0.15 to 0.64); and *social comparison* combined with *prompts and cues* (0.33, 95%CI 0.22 to 0.44).

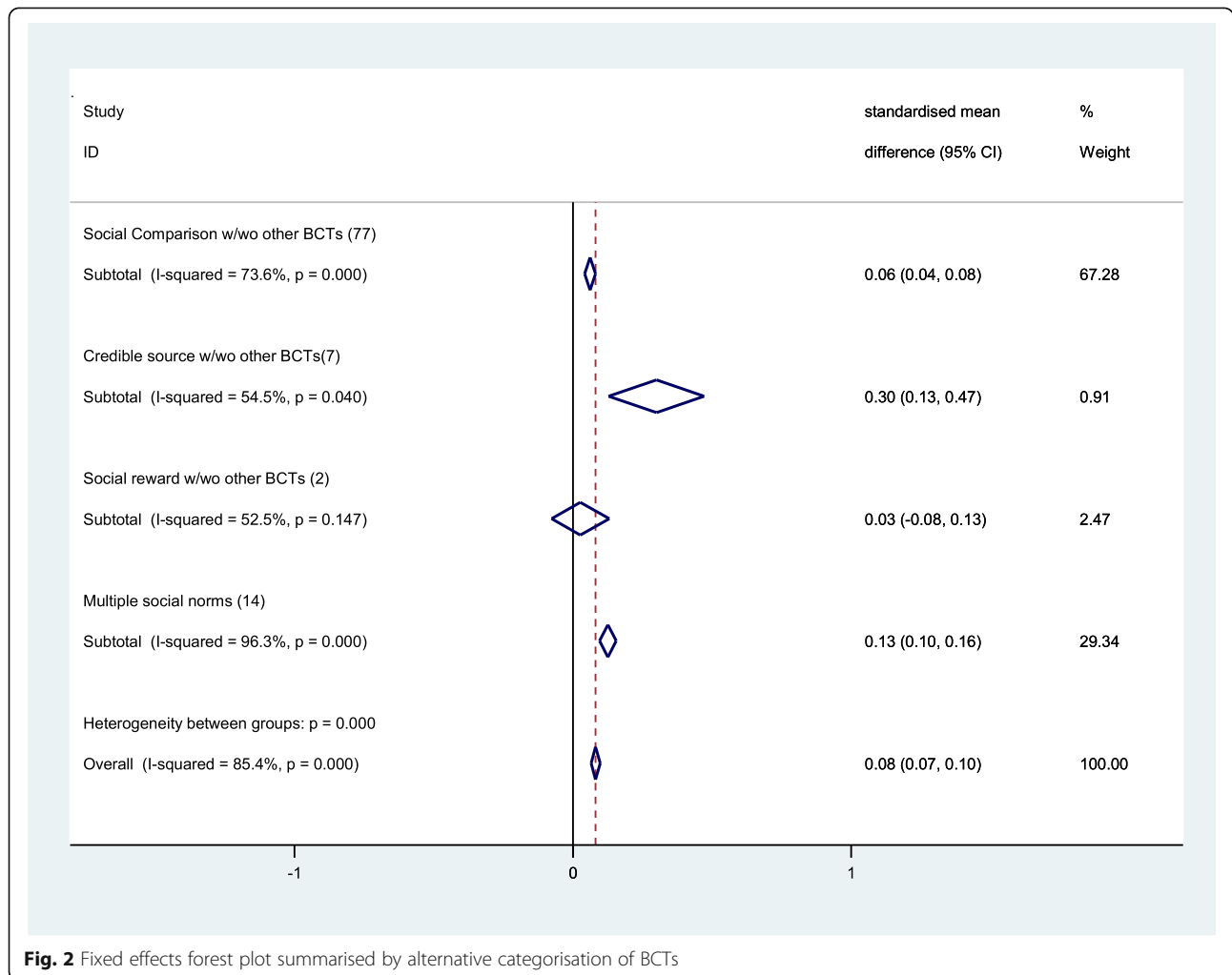


Fig. 2 Fixed effects forest plot summarised by alternative categorisation of BCTs

Table 4 Intervention effects calculated from meta-analysis and network meta-analysis, ordered by effect size (intervention v control)

Type of social norms intervention	Number of comparisons for meta-analysis (network meta-analysis)	SMD meta-analysis (95%CI) $n = 100$	SMD network meta-analysis (95%CI) $n = 102$	Probability of being the best intervention (%)
Social comparison + social reward	2	0.39(0.15 to 0.64)	0.39 (0.15 to 0.64)	59.2
Social comparison + prompts/cues	5	0.33(0.22 to 0.24)	0.33 (0.22 to 0.44)	22.2
Credible source ^a	7	0.30(0.13 to 0.47)	0.30 (0.13 to 0.47)	18.6
Social comparison + credible source ^a	8(10)	0.16(0.12 to 0.19)	0.16(0.12 to 0.20)	0.0
Social comparison + social support (unspecified)	7	0.10(0.04 to 0.16)	0.10 (0.04 to 0.16)	0.0
Other multiple social norms BCTs	4	0.07(0.03 to 0.12)	0.07(0.03 to 0.12)	0.0
Social comparison	33(35)	0.05(0.03 to 0.08)	0.05(0.03 to 0.08)	0.0
Social comparison + other BCTs	23	0.04(0.00 to 0.08)	0.04(0.00 to 0.08)	0.0
Social reward	2	0.03(- 0.08 to 0.13)	0.03(- 0.08 to 0.13)	0.0
Social comparison + instructions on how to perform the behaviour + prompts/cues	5	0.01(- 0.10 to 0.11)	0.01(- 0.10 to 0.11)	0.0
Social comparison + info on health consequences	4	- 0.14 (- 0.33 to 0.05)	- 0.14(- 0.33 to 0.05)	0.0

^aWith/without other BCTs

Social comparison delivered with *credible source* (0.16, 95%CI 0.12 to 0.19), on its own (0.05, 95%CI 0.03 to 0.08) or with *social support (unspecified)* (SMD 0.10, 95%CI 0.04 to 0.16) were all effective, on average, compared to control. This was confirmed by network meta-analysis. Table 5 shows the different contexts and settings for the social norms BCTs and there does not appear to be any obvious patterns of use of the BCTs in particular contexts: social comparison, credible source and social reward are each used in multiple different contexts either alone or alongside other BCTs. Regression analysis suggests that results were consistent even after adjustment for context and setting. Illustrative case studies providing examples of the three intervention types found to be most effective (credible source, social comparison with prompts/cues, social comparison and social reward) are shown in Table 6.

Context and mode of delivery

Meta-analysis suggested that social norms interventions were effective in a variety of different contexts. The effect was seen with doctors on average (SMD 0.08, 95%CI 0.07 to 0.10, $n = 68$) and other healthcare workers (SMD 0.08, 95%CI 0.04 to 0.12, $n = 12$), but not with nurses and allied healthcare workers (SMD - 0.01, 95%CI - .012 to 0.11, $n = 5$). They appeared successful across a range of clinical behaviours, including prescribing (SMD 0.11, 95%CI 0.09 to 0.13, $n = 21$), arranging, conducting or administering tests/assessments (SMD 0.10, 95%CI 0.06 to 0.13, $n = 21$), and management and communication around health conditions (SMD 0.06, 95%CI 0.01 to 0.12, $n = 23$), but may be less effective with handwashing

(SMD 0.04, 95%CI - 0.05 to 0.13, $n = 3$) and referrals to other health services (SMD - 0.08, 95%CI - 0.23 to 0.07, $n = 3$). The effects were similar in primary (SMD 0.07, 95%CI 0.05 to 0.09, $n = 56$) and secondary care (SMD 0.12, 95%CI 0.07 to 0.18, $n = 27$) but may be less effective in community (SMD 0.02, 95%CI - 0.05 to 0.10, $n = 4$) and care home (SMD 0.03, 95%CI - 0.05 to 0.10, $n = 4$) settings. The effect appears to be consistent, regardless of whether a peer benchmark (0.06, 95%CI 0.02 to .011, $n = 13$) or the average (0.11, 95%CI 0.09 to 0.13, $n = 67$) is included. On average, they were slightly less effective in increasing behaviours (e.g. increasing diabetes testing) than at reducing behaviours (e.g. reducing antibiotic prescriptions). The effect was similar regardless of who delivered the intervention and whether it came from within the organisation or from an external source. Interventions that were delivered once (0.25, 95%CI 0.21 to 0.30, $n = 28$) were more effective than those delivered more frequently (0.06, 95%CI 0.04 to 0.08, $n = 47$). Delivery by website was most effective (0.23, 95%CI 0.15 to 0.31, $n = 8$); delivery by email, in writing, and in mixed format were all consistent with the average effect, but face-to-face appeared to be ineffective (- 0.01, 95%CI - 0.06 to 0.03, $n = 14$). The number of studies in some of these categories was low (nurses and allied healthcare workers, handwashing, referrals to other services, community and care homes), and none of the pre-planned covariates for context and setting appeared to explain much of the heterogeneity in meta-regression, suggesting that any conclusions about context and mode of delivery should remain cautious.

Table 5 Key trial characteristics by type of comparisons

Type of comparison	Test of SC	Test of CS	Test of SR	Test of SC + SR	Test of SC + social support (unspecified)	Test of SC + support + prompts and cues	Test of SC + prompts and cues	Test of SC + info on health consequences	Test of SC + health instructions + prompts/ cues	Test of SC + others BCTs	Test of CS + other BCTs	Test of SC + SR + other BCTs	Test of multiple SNs + other BCTs
	33	3	1	2	2	7	5	4	5	25	4	4	1
Number of comparisons with primary outcome data													
Target Behaviour													
Prescribing	15(45%)	1(100%)	1(50%)	2(100%)	2(29%)	1(20%)	3(75%)	1(20%)	11(44%)	1(25%)	1(25%)	1(100%)	1(25%)
Hand/hygiene	7(21%)				1(14%)	3(60%)	1(25%)	3(60%)	4(16%)	1(25%)	1(25%)	1(25%)	1(25%)
Tests													
Referrals	5(15%)	3(100%)	1(50%)	2(29%)	12(14%)	1(14%)	2(29%)	1(20%)	5(20%)	2(8%)	1(25%)	1(25%)	2(50%)
Man/comm	6(18%)				1(14%)	1(20%)			2(8%)				
Other													
Multiple													
Type of HCP													
Doctor GP	16(48%)				4(57%)	2(40%)	1(25%)	4(80%)	11(44%)	2(50%)	1(25%)	1(100%)	1(25%)
Doctor secondary	4(12%)	3(100%)	1(50%)	2(100%)	1(14%)	1(20%)	1(20%)	1(20%)	2(12%)	1(25%)	1(25%)	1(100%)	1(25%)
Other HCP	4(12%)		1(100%)						1(4%)	0(0%)	0(0%)	1(100%)	1(25%)
Mixed/team	9(27%)				2(29%)	2(40%)	2(40%)	1(20%)	10(40%)	2(50%)	2(50%)	1(100%)	1(25%)
Setting													
Primary	18(55%)				4(57%)	4(80%)	1(25%)	5(100%)	17(68%)	2(50%)	1(25%)	1(100%)	1(25%)
Hospital	6(18%)	3(100%)		2(29%)	2(29%)	1(20%)	2(50%)	2(50%)	6(24%)	2(50%)	2(50%)	1(100%)	2(50%)
Community	1(3%)		1(100%)	1(50%)	1(14%)				1(4%)				1(25%)
Care/nursing	0(0%)												
Mixed	7(21%)												
Other	1(3%)												

SC social comparison, CS credible source, SR social reward

Table 6 Case studies—summary descriptions of interventions for example studies of the three intervention types found to be most effective

Study Trial design Target healthcare worker	Aims	Outcome measure SMD(95% CI)	Control arm	Intervention description
Credible source + social comparison				
Hallsworth et al. (2016) [32] RCT Doctor (primary care)	To reduce the number of unnecessary prescriptions of antibiotics by GPs in England	The rate of antibiotic items dispensed per 1000 population 0.13 (0.03 to 0.29)	Delayed intervention (after the end of the trial (no BCTs were coded).	A letter was sent to GPs from the Chief Medical Officer. The letter stated that the practice was prescribing antibiotics at a higher rate than 80% of practices in its NHS Local Area Team, and used three concepts from the behavioural sciences. The first was social norm information about how the recipient's practices prescribing rate compared with other practices in the local area. Second, the letter was addressed from a high-profile figure with the assumption that this would increase the credibility of its content. Finally, the letter presented three specific, feasible actions that the recipient could do to reduce unnecessary prescriptions of antibiotics: giving patients advice on self-care, offering a delayed prescription and talking about the issue with other prescribers in his or her practice. The letter was accompanied by a copy of the patient-focused "Treating your infection" leaflet, which acted to reinforce the message of the letter by supporting delayed or reduced prescribing. (9.1 Credible source; 6.2 Social comparison, 2.2 Feedback on behaviour, 4.1 Instruction on how to perform the behaviour).
Social comparison + prompts/cues				
Vellinga et al (2016) [33] Cluster RCT Doctor-GP	To increase the number of first-line antimicrobial prescriptions for suspected urinary tract infections (UTIs) in adult patients	Adherence to guidelines for antimicrobial prescribing in primary care 0.55 (0.32 to 0.77)	Phase 1—a coding workshop: routine coding for UTIs using standardised codes were demonstrated. The purpose of this was to facilitate the generation of electronic audit and feedback reports (not available to control until after the trial). Control practices then provided 'usual care' for the remainder of the intervention (no BCTs were coded).	Arm A: phase 1—a coding workshop (same as control). Phase 2—interactive workshops were designed to promote changes in antimicrobial prescribing for the treatment of UTIs by presenting an overview of prescribing and antimicrobial resistance, discussing the role of the GP in the spread of anti-microbial resistance. A computer prompt was developed for use within the selected GP practice management software system. This prompt summarised the recommendations for first-line antimicrobial treatment and appeared on the computer screen when the GP entered the International Classification of Primary Care code (U71) for 'cystitis, urinary infection, other'. This prompt also reminded the GP to collect patients' mobile telephone numbers. Electronic audit and feedback reports were available to download by GPs. These reports provided the practice with information on antimicrobial prescribing for UTI in comparison with the aggregated information from the other

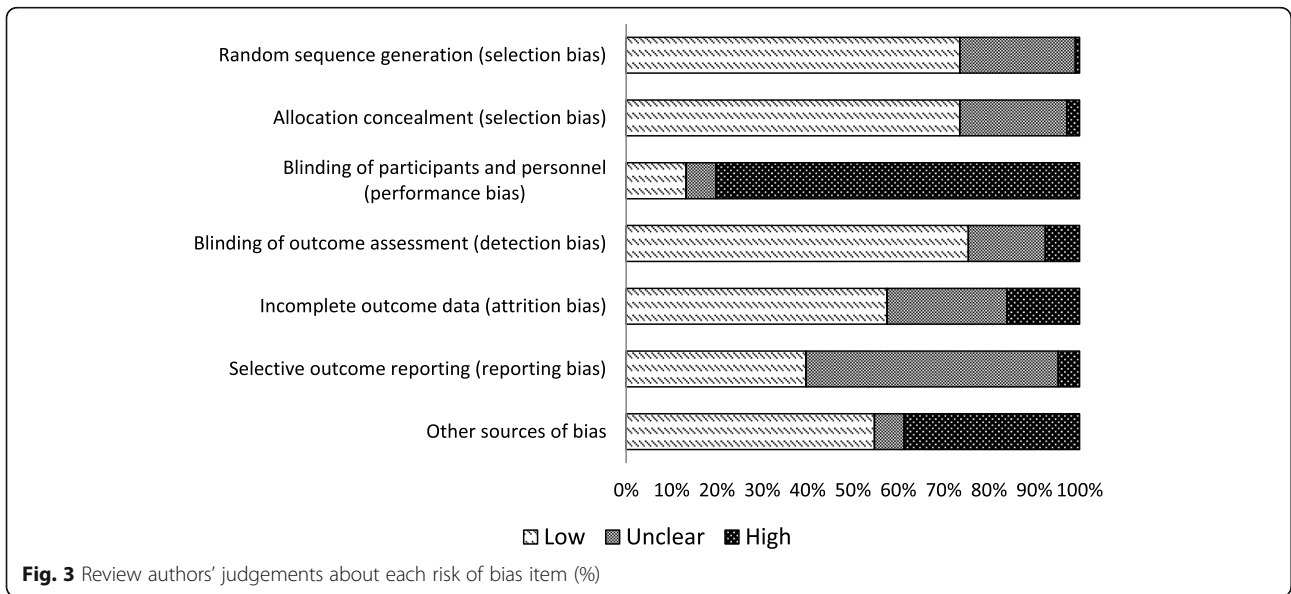


Fig. 3 Review authors' judgements about each risk of bias item (%)

Risk of bias

A summary of each risk of bias item across the studies is shown in Fig. 3. Risk of bias was high in 80% of trials for the blinding of participants and personnel domain and so we cannot rule out the possibility of response bias. This high risk of bias was mainly due to the nature of the interventions (i.e. many of the studies were cluster trials, randomised at the hospital or clinic level, making blinding impractical). In a sensitivity analysis restricting the meta-analysis to trials at low risk of bias for each key domain, the overall treatment effect changed little, suggesting the results were robust. There were five studies at high risk of bias for outcome reporting and 59 with

unclear risk of bias. A funnel plot (Fig. 4) identified that the review may be missing some unpublished negative trials, or including more positive trials than expected, suggesting selective outcome reporting.

Discussion

Summary of evidence

Social norms interventions can be an effective approach to changing the clinical behaviours of healthcare workers. Meta-analysis showed social norms interventions were associated with an improvement in healthcare worker clinical behaviour outcomes of 0.08 SMDs (95%CI 0.07 to 0.010, $n = 100$ comparison) and an

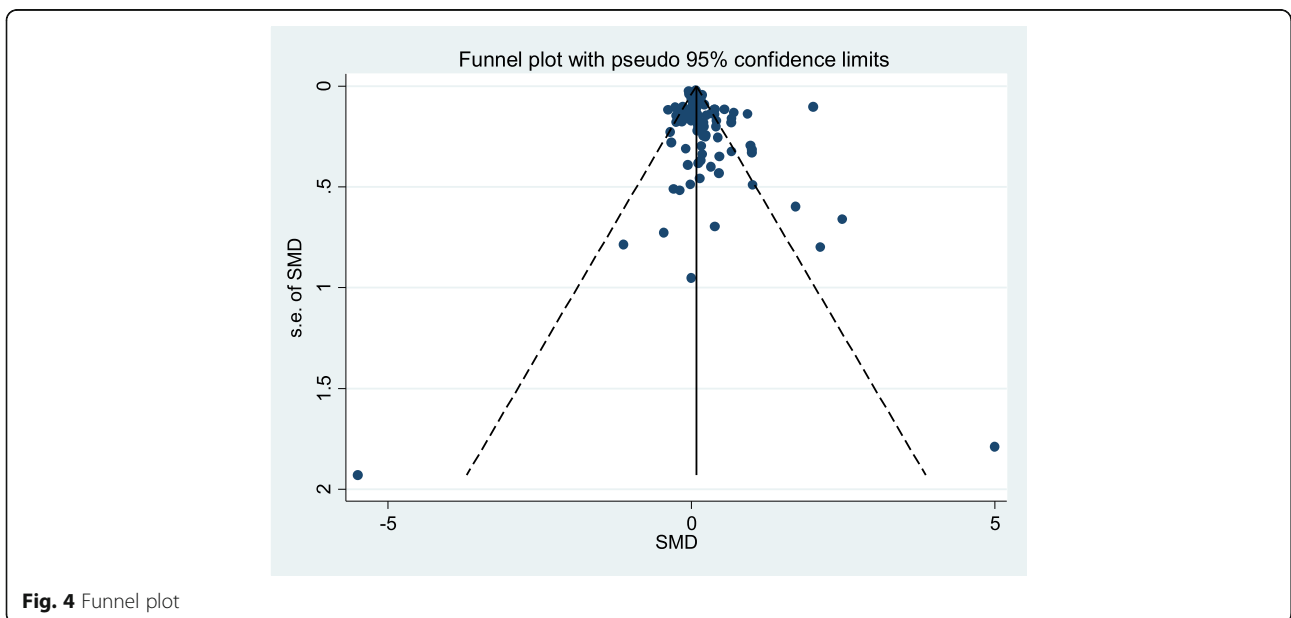


Fig. 4 Funnel plot

improvement in patient health outcomes of 0.17 SMD (95%CI 0.14 to 0.20, $n = 14$ comparisons), on average.

There was a large amount of heterogeneity, with some studies reporting substantially higher or lower effects. There was strong evidence from multiple studies that interventions involving *social comparison* or *credible source*, with and without other BCTs, were effective on average, both separately and together. *Social comparison* is effective when combined with various other BCTs including *social support (unspecified)* but it appears to be most effective when combined with *prompts/cues*. *Social reward* appeared not to be effective when used alone but had an above-average effect when combined with *social comparison*. The effect of social norms interventions remained clear in the meta-regression, even after taking into account context and setting.

Meta-analyses exploring context and delivery showed that social norms were effective with a variety of healthcare workers, in primary and secondary care, and across a range of clinical behaviours. On average, social norms interventions were more effective for reducing than increasing behaviours. Interventions appeared equally effective regardless of whether they came from an internal or external source. In contrast to previous studies [14], delivering the intervention once appeared to be more effective than frequent delivery: one explanation for this, which warrants further investigation, is whether frequent delivery is associated with attempts to change intractable behaviours.

Sensitivity analyses found the overall treatment effect to be robust and not strongly influenced by trials which scored high/unclear risk of bias across key domains. There is a possibility of response bias due to lack of blinding. While it is difficult to blind healthcare workers in these trials, there were examples where the risk of response bias was minimised, e.g. cluster trials where the healthcare worker was not informed of the existence of the trial.

Discussion of findings in relation to the literature

A Cochrane systematic review ($n = 140$) of the effect of A&F on healthcare worker behaviour and patient health outcomes [14] found a wide variation in the effect of A&F and recommended future research to explore how this variation, related to the intervention design, context and recipient [11]. The results of our review contribute to this agenda by suggesting aspects of the design of A&F interventions that are associated with positive outcomes: (1) highlighting that a *credible source* approves of the desired behaviour; (2) feedback on an individual's behaviour is likely to be more effective if accompanied by *social comparison*; (3) complex interventions involving multiple social norms seem to be effective; (4) *social comparison* seems to be enhanced by the use of *prompts and cues*,

such as computerized pop-ups recommending actions to GPs when particular symptoms or diagnoses are entered into an electronic system [35], but the benefit of *prompts and cues* may only hold when the healthcare worker understands how to do the behaviour. The effects of social norms were reasonably consistent across a range of healthcare workers, behaviours and settings. In contrast to an earlier review of A&F [14], delivering the intervention once appeared to be sufficient and sending the intervention by website or other computerised format was most effective. Our results align with findings from a recent synthesis of qualitative literature on A&F which found that letting healthcare workers know how their performance relates to that of their peers (*social comparison*) and providing opportunities for peer discussion (*social support (unspecified)*) were valuable in changing behaviour [6]. However, our finding that face-to-face interventions were less effective than remotely delivered interventions contrasts with results for meta-analyses of smoking cessation interventions where personalised interventions were associated with greater effectiveness [36]. Recent literature suggests that de-implementation is often even more challenging than implementation due to a number of psychological biases: health professionals tend to focus on information that confirms their established beliefs; people are more concerned about losses than gains; and a sense of professional autonomy strengthens attachment to established practices [37, 38]. Given the challenges of de-implementation our finding that social norms interventions were more effective in increasing behaviour than decreasing it are perhaps not surprising.

A recent overview of 67 systematic reviews on promoting professional behaviour change in healthcare found that the most effective interventions were educational outreach using academic detailing, A&F and reminders [39]. Using normalization process theory as a theoretical lens, the authors concluded that interventions that seek to 'restructure and reinforce new practice norms' (opinion leaders, educational meetings and materials/guidelines) and those which 'associate practice norms with peer and reference group behaviours' (including A&F and academic detailing, where a target healthcare worker receives individual support or advice from someone else with expertise in that area) are most likely to be successful in changing clinical behaviour. Combining the two approaches together may be particularly effective, by creating clear rules of conduct and encouraging individuals to follow their peers [39]. Interventions that seek to change attitudes were less likely to be successful. The importance given to peer and reference group behaviours in this previous study justifies our efforts to identify which social norms interventions are associated with success.

The effect sizes seen in this review appear to be similar to other reviews of interventions to change health

professional behaviour [40]. Baskerville et al found that practice facilitation was associated with an improvement of 0.56 SMD (95% CI 0.43 to 0.68) in guideline adoption in primary care. Baker [41] reported that tailored interventions to overcome barriers to change are associated with an odds ratio for the improvement in professional behaviour of 1.51 (95% CI 1.16 to 2.01) which corresponds to an SMD of approximately 0.24 (95% CI 0.09 to 0.39). The modest effects size seen for social comparison appears in line with that observed by Ivers who found that Audit and feedback improved binary behavioural outcomes by a median of 4.3 percentage points and continuous outcomes by a median of 1.3 percentage points. In a meta-synthesis of systematic reviews of health behaviour change in general, Johnson found effect sizes between 0.08 to 0.45 [42].

Strengths and limitations

Our search strategy was developed through an iterative process, with input from an Information Scientist. However, it is possible that the strategy may have missed some relevant interventions if social norms BCTs or behaviour change theories were not mentioned in the title or abstract.

We included studies regardless of outcome measure, and we converted any available outcome into a standardized mean difference: this meant we were able to summarise all the available evidence in one analysis. The included trials incorporated a variety of contexts and settings; trial designs and units of analysis. This has led to a heterogeneous review; and we acknowledge the limitations of this approach. The magnitude of effects for the most promising behaviour change interventions were around 0.3 SMDs, which relative to the between study variability $\tau(0.2)$ does seem to indicate an important effect.

Trials were excluded from the review where the intervention did not target a specific behaviour: for example, if the intervention was aimed at a healthcare worker with the intention of reducing patient blood pressure, but did not make explicit what behaviour(s) were expected of the healthcare worker to achieve the reduction. These exclusions occurred because, if a behaviour is not specified, it is not possible to determine whether or not an intervention actually targeted that behaviour and change in that behaviour (our primary outcome) cannot be assessed. This approach is consistent with the coding instructions of the BCTTv1 [16]. There is a potential risk that we have excluded some studies where there was a target behaviour but it was poorly reported.

We used the BCTTv1 [16], which has been based on a significant body of research, to code for BCTs that could be associated with the effectiveness of interventions. However, BCTs were only coded based on published reports and we did not ask study authors for intervention manuals due to time constraints. Therefore, it is possible that our

coding did not represent all actual BCTs as designed and delivered. The authors of the BCTTv1 have also acknowledged that extension or modification of the BCTTv1 could be appropriate in the future. It is therefore possible that some BCTs that do not yet feature in the BCTTv1 could have been presented alongside social norms BCTs and were missed during the BCT coding exercise.

Ten small studies without suitable outcome measures were omitted from the meta-analysis and some missing information (such as ICCs and standard deviations) were imputed, but sensitivity analyses suggested no significant impact on the review.

The primary approach to meta-analysis was fixed effects [43], which summarises the evidence in these trials, rather than estimating a common underlying treatment effect [44]. This topic is highly contested, so random effects was also undertaken for the most important analyses, as planned. In all analyses the fixed and random effects approaches produced a result in the same direction, with the random effects approach resulting in a higher effect for the intervention because it gives greater weight to smaller studies. The conclusions of the review would be similar, regardless of whether fixed or random effects were used.

All of the meta-analysis was undertaken on the basis of comparisons: the BCTs in the control arm were subtracted from those in the intervention arm to capture BCTs that were actively tested in each study. The active ingredient was what is left of the intervention when the control arm is taken away. This is a suitable approach to examining the effect of the various social norm BCTs, but a limitation is that some interaction effects may have been missed.

The asymmetry of the funnel plot suggested that the review may have missed some unpublished negative trials or be at risk of bias from selective outcome reporting. The resources were not available for translation or to request unpublished material from authors of included studies, so some relevant studies may have been omitted. A single behaviour outcome was selected from every trial using published reports which may have put the review at risk from selective outcome reporting; priority was given to the pre-specified primary outcome. Sensitivity analysis including only those trials with either a relevant pre-specified primary outcome or single relevant behavioural outcome suggested that results were robust to selective outcome reporting.

Further research

Credible source has been identified as an effective intervention component. Yet, it is not commonly used in the health setting to change the behaviour of healthcare workers (only 18% of the comparisons identified in the present review). This may be due to *credible source* lacking formal conceptualisation in the health setting so,

whilst it may be used in practice, it is not well-reported. Additional work is needed to develop *credible source* interventions for use in the NHS, such as, whom the target audience would consider as *credible sources*: for example, seniority may not necessarily be perceived as the same as credible. As a first step, a narrative synthesis of the trials using *credible source* in this review, together with the qualitative papers, process evaluations and protocols associated with those trials, would provide further insights into the *credible source* interventions that are associated with more successful outcomes. Qualitative work with healthcare workers, managers and policy-makers is also needed to understand the acceptability and feasibility of *credible source*, *social comparison* and *social reward* interventions and to understand who the most credible sources are.

Social comparison is currently used more frequently with healthcare workers than *credible source*. We identified a high level of heterogeneity in the effectiveness of *social comparison*. We have started to unravel this heterogeneity, and this research suggests that *social comparison* can successfully be enhanced by the addition of *social reward*, *prompts and cues* or *social support (unspecified)*; but further research is warranted. The heterogeneity could potentially be explained by differences in how social comparisons are facilitated and what kind of comparisons are made, and not simply by the combination of BCTs it is delivered with or without. For example, social comparisons may have a different effect depending on the reference frame (e.g. whether one identifies with those compared to) or depending on the direction of the comparison (i.e. upward or downward comparison). Further investigation into the factors that moderate the effect of social comparison is warranted.

The methodological quality of trials was mixed. The review included some large factorial trials that tested several behaviour change interventions simultaneously, which can be an efficient design for exploring different components of behaviour change interventions and their interactions. Multiphase Optimization Strategy may be a useful framework that can be applied to factorial designs for identifying which combination and sequence of components (e.g. BCTs and mode of delivery) can produce optimal outcomes [45]. Some trials also used novel methods to minimize bias such as ‘attention’ controls where participants were given the identical behaviour change intervention for an alternative target behaviour: this type of design is to be encouraged.

Conclusions

Social norms interventions are an effective method of changing healthcare worker clinical behaviour. Although the overall result is modest and very variable, there is the potential for social norms interventions to be applied at scale and

have a significant effect on clinical behaviour and resulting patient health outcomes. Both *credible source* and *social comparison* were effective. *Social comparison* was particularly effective when combined with *prompts and cues*. These interventions were found to be effective in a variety of NHS contexts and across a range of modes of delivery.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13012-020-01072-1>.

Additional file 1: Appendix 1. Search Strategy. **Appendix 2.** Data Extraction Form. **Appendix 3.** Behaviour Change Techniques (BCTs) Extraction Form. **Appendix 4.** Inter-Rater Agreement for BCT Coding. **Appendix 5.** Included Study References. **Appendix 6.** Study and Intervention Characteristics of Included Comparisons. **Appendix 7.** Sensitivity Analyses.

Abbreviations

A&F: Audit and Feedback; BCT: Behaviour Change Technique; BCTTV1: Behaviour Change Techniques Taxonomy Version 1; BNI: British Nursing Index; CI: Confidence Interval; CINAHL: Cumulative Index of Nursing and Allied Health Literature; Cochrane CENTRAL: Cochrane Central Register of Controlled Trials; EMBASE: Excerpta Medica dataBASE; GP: General practitioner; ICC: Intra-class correlation coefficient; MEDLINE: Medical Literature Analysis and Retrieval System Online; MeSH: Medical Subject Headings; PABAK: Prevalence-Adjusted and Bias-Adjusted Kappa; PRISMA: Preferred Reporting Items for Systematic Reviews; PROSPERO: The International Prospective Register of Systematic Reviews; RCT: Randomised controlled trial; SMD: Standard mean difference; SN: Social norm

Acknowledgements

We thank the independent members of the Study Steering Committee: Robbie Foy (Chair), Professor of Primary Care, University of Leeds; Marie Johnston, Professor of Health Psychology, University of Aberdeen, Sofia Dias, Professor in Health Technology Assessment, University of York, and Manoj Mistry, lay member. They were very generous with their time and provided encouragement and wise counsel throughout the project. Marie Johnston came all the way from Aberdeen to run a training workshop on BCT coding, and we are also grateful to her for providing comments on an earlier draft of this manuscript. Thank you to Jane Roberts for writing and conducting the searches and double-coding the BCTs. We are grateful to Jack Wilkinson who was employed as a researcher in the early stages of the review until he was successful in winning a University of Manchester presidential fellowship.

Authors' contributions

MYT (Research Associate, Health Psychology) wrote the first draft; all authors commented on drafts and approved the manuscript. SC (Senior Lecturer, Health Services Research and Statistics) conceived of the idea for the review and managed the project. SC, RP (Senior Lecturer, Health Psychology), SR, BB (Senior Academic GP and Honorary Consultant) and MYT contributed to the protocol. RP was the lead for health psychology, SR was the lead statistician and BB was the clinical lead. MYT, SC and SR were involved in the screening of studies. MYT, SC, SR and EH (Research Associate, Statistics) did the data extraction. MYT, SC and RP (with Jane Roberts) did the BCT coding. SR (Senior Research Fellow, Statistics) converted all the outcome measures to SMDs, conducted the meta-analysis, meta-regression and network meta-analysis of the healthcare worker behaviour outcomes and prepared the results for publication. EH conducted the meta-analysis of patient health outcomes. MYT, LM (PhD student, Psychology) and SC helped to prepare the data for analysis and undertook other descriptive analysis. The author(s) read and approved the final manuscript.

Funding

This project is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research, reference 17/06/06—The impact of social norms interventions on clinical behaviour change among healthcare workers: a systematic review. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester M13 9PL, UK. ²National Institute of Health Research Behavioural Science Policy Research Unit, Population Health Sciences, Baddiley-Clark Building, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne NE2 4AX, UK. ³Manchester Centre for Health Psychology, Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester M13 9PL, UK. ⁴Health e-Research Centre, Farr Institute for Health Informatics Research, Faculty of Biology Medicine and Health, University of Manchester, Manchester M13 9PL, UK. ⁵Centre for Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Manchester M13 9PL, UK.

Received: 12 May 2020 Accepted: 9 December 2020

Published online: 07 January 2021

References

- Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ*. 2004;329(7456):15–9.
- Howard RL, Avery AJ, Slavenburg S, Royal S, Pipe G, Lucassen P, et al. Which drugs cause preventable admissions to hospital? A systematic review. *British journal of clinical pharmacology*. 2007;63(2):136–47.
- Courtenay M, Rowbotham S, Lim R, Peters S, Yates K, Chater A. Examining influences on antibiotic prescribing by nurse and pharmacist prescribers: a qualitative study using the Theoretical Domains Framework and COM-B. *BMJ Open*. 2019;9(6):e029177.
- Paluck ELB, L. Social norms marketing aimed at gender based violence: a literature review. New York: International Rescue Committee; 2010.
- Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V. The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science*. 2007;18(5):429–34.
- Brown B, Gude WT, Blakeman T, van der Veer SN, Ivers N, Francis JJ, et al. Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implement Sci*. 2019;14(1):40.
- Festinger L. A Theory of Social Comparison Processes. *Human Relations*. 1954;7(2):117–40.
- Tajfel H, Turner JC. The social identity theory of inter-group behavior. In: Worchel S, Austin WG, editor. *Psychology of Intergroup Relations*. Chicago: Nelson-Hall; 1986.
- Ajzen I. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*. 1991;50(2):179–211.
- Cialdini RB, Kallgren CA, Reno RR. A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*. 1991;24:201–34.
- Ivers NM, Sales A, Colquhoun H, et al. No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention. *Implement Sci*. 2014;9:14. <https://doi.org/10.1186/1748-5908-9-14>.
- Carney PA, Abraham L, Cook A, Feig SA, Sickles EA, Miglioretti DL, et al. Impact of an educational intervention designed to reduce unnecessary recall during screening mammography. *Academic Radiology*. 2012;19(9):1114–20.
- Ivers NM, Tu K, Young J, Francis JJ, Barnsley J, Shah BR, et al. Feedback GAP: pragmatic, cluster-randomized trial of goal setting and action plans to increase the effectiveness of audit and feedback interventions in primary care. *Implement Sci*. 2013;8:142.
- Ivers N, Jamtvedt G, Flottorp S, Young JM, Odaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev*. 2012;6:CD000259.
- Gardner B, Whittington C, McAteer J, Eccles MP, Michie S. Using theory to synthesise evidence from behaviour change interventions: the example of audit and feedback. *Social science & medicine*. 2010;70(10):1618–25.
- Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med*. 2013;46(1):81–95.
- Cotterill S, Powell R, Rhodes S, Brown B, Roberts J, Tang MY, et al. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. *Syst Rev*. 2019;8(1):176.
- Deeks JJ, Higgins JPT, Altman DG. Chapter 10: Analysing data and undertaking meta-analyses. In: *Cochrane Handbook for Systematic Reviews of Interventions* version 60 (updated July 2019); 2019.
- Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*. 2002;21:1559–73.
- Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PLoS One*. 2013;8(10):e76654.
- Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6(7):e1000097.
- Covidence. Covidence: better systematic review management 2019 [Available from: www.covidence.org].
- Effect Practice and Organisation of Care. Data collection form: EPOC resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services; 2013. [Available from: <https://epoc.cochrane.org/resources/epoc-resources-review-authors>].
- Abraham C, Wood CE, Johnston M, Francis J, Hardeman W, Richardson M, et al. Reliability of Identification of Behavior Change Techniques in Intervention Descriptions. *Annals of Behavioral Medicine*. 2015;49(6):885–900.
- Chen G, Faris P, Hemmelgarn B, Walker RL, Quan H. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Medical Research Methodology*. 2009;9:5.
- Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ*. 2019;364:k4817.
- Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*. 2000;19:3127–31.
- The Cochrane Collaboration 16.3.6. Approximate analyses of cluster-randomized trials for meta-analysis: inflating standard errors. In: Higgins JPT, Green S, editor. *Cochrane Handbook for Systematic Reviews of Interventions (Version 510)* 2011.
- Harris RB, M.; Deeks, J.; Harbord, R.; Altman, D.; Steichen, T.; Sterne, J. METAN: Stata module for fixed and random effects meta-analysis. *Statistical Software Components*. 2006;S456798.
- Harbord RMH, J.P.T. Meta-Regression in Stata. *The Stata Journal*. 2008;8(4): 493–519.
- Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Med Res Methodol*. 2008;8:79.
- Hallsworth M, Chadborn T, Sallis A, Sanders M, Berry D, Greaves F, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. *The Lancet*. 2016;387(10029):1743–52.
- Vellinga A, Galvin S, Duane S, Callan A, Bennett K, Cormican M, et al. Intervention to improve the quality of antimicrobial prescribing for urinary tract infection: a cluster randomized trial. *Canadian Medical Association Journal*. 2016;188(2):108–15.
- Persell SD, Doctor JN, Friedberg MW, Meeker D, Friesema E, Cooper A, et al. Behavioral interventions to reduce inappropriate antibiotic prescribing: a randomized pilot trial. *BMC Infectious Diseases*. 2016;16(1):373.
- Duane S, Callan A, Galvin S, Murphy AW, Domegan C, O'Shea E, Cormican M, Bennett K, O'Donnell M, Vellinga A. Supporting the improvement and management of prescribing for urinary tract infections (SIMPLE): protocol for a cluster randomized trial. *Trials*. 2013;14(441):1–13.
- Black N, Eisma MC, Viechtbauer W, Johnston M, West R, Hartmann-Boyce J, Michie S, de Bruin M. Variability and effectiveness of comparator group

- interventions in smoking cessation trials: a systematic review and meta-analysis. *Addiction*. 2020;115(9):1607-17.
37. van Bodegom LD, Davidoff F, Marang-van de Mheen PJ. Implementation and de-implementation: two sides of the same coin? *BMJ Quality & Safety*. 2017;2(6):495-501.
 38. Ubel PA, Asch DA. Creating value in health by understanding and overcoming resistance to de-innovation. *Health Affairs*. 2015;34(2):239-44.
 39. Johnson MJ, May CR. Promoting professional behaviour change in healthcare: what interventions work, and why? A theory-led overview of systematic reviews. *BMJ Open*. 2015;5(9):e008592.
 40. Baskerville NB, Liddy C, Hogg W. Systematic review and meta-analysis of practice facilitation within primary care settings. *Annals of Family Medicine*. 2012;10(1):63-74.
 41. Baker R, Camosso-Stefinovic J, Gillies C, Shaw EJ, Cheater F, Flottorp S, Robertson N. Tailored interventions to overcome identified barriers to change: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev*. 2010;3:CD005470.
 42. Johnson BT, Scott-Sheldon LAJ, Carey MP. Meta-Synthesis of Health Behavior Change Meta-Analyses. *American Journal of Public Health*. 2010;100(11):2193-8.
 43. Poole CG, S. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology*. 1999;150(5):169-75.
 44. Higgins JPT, Lopez-Lopez JA, Becker BJ, Davies SR, Dawson S, Grimshaw JM, et al. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Glob Health*. 2019;4(Suppl 1):e000858.
 45. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. *Am J Prev Med*. 2007;32(5 Suppl):S112-8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



6.7 PAPER 7

Rhodes S, Dias S, Wilkinson J, Cotterill S. Synthesis of data from trials of interventions designed to change health behaviour; a case study. Authorea. 2021.

<https://doi.org/10.22541/au.163256037.77153698/v1>

Synthesis of data from trials of interventions designed to change health behaviour; a case study

Sarah Ann Rhodes¹, Sofia Dias², Jack Wilkinson¹, and Sarah Cotterill¹

¹The University of Manchester

²University of York

September 25, 2021

Abstract

Many complex healthcare interventions aim to change the behaviour of patients or health professionals, e.g. stopping smoking or prescribing fewer antibiotics. This prompts the question of which behaviour change interventions are most effective. Synthesising evidence on the effectiveness of a particular type of behaviour change intervention can be challenging because of the high levels of heterogeneity in trial design. Here we use data from a published systematic review as a case study and compare alternative methods to address this heterogeneity. One important source of heterogeneity is that compliance to a desired behaviour can be measured and reported in a variety of different ways. In addition, interventions designed to target behaviour can be implemented at either an individual or group level leading to trials with varying layers of clustering. To handle heterogeneous outcomes we can either convert all effect estimates to a common scale (e.g. using standardised mean differences) or have separate meta-analyses for different types of outcome measure (binary and continuous measures). To address the clustering structure, adjusted standard errors can be used with the inverse variance method, or weights can be assigned based on a consistent level of clustering, such as the number of healthcare professionals. A graphical method, the albatross plot utilises reported p-values only, and can synthesise data with both heterogeneous outcomes and clustering with minimal assumption and data manipulation. Based on these methods, we reanalysed our data in four different ways and have discussed the strengths and weaknesses of each approach.

Synthesis of data from trials of interventions designed to change health behaviour; a case study.

Sarah Rhodes¹

Email: sarah.a.rhodes@manchester.ac.uk

Sofia Dias²

Email: sofia.dias@york.ac.uk

Jack Wilkinson¹

Email: jack.wilkinson@manchester.ac.uk

Sarah Cotterill¹

Email: sarah.cotterill@manchester.ac.uk

¹Centre for Biostatistics, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology Medicine and Health, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

²Centre for Reviews and Dissemination, University of York, York, YO10 5DD, UK

Abstract

Many complex healthcare interventions aim to change the behaviour of patients or health professionals, e.g. stopping smoking or prescribing fewer antibiotics. This prompts the question of which behaviour change interventions are most effective. Synthesising evidence on the effectiveness of a particular type of behaviour change intervention can be challenging because of the high levels of heterogeneity in trial design. Here we use data from a published systematic review as a case study and compare alternative methods to address this heterogeneity. One important source of heterogeneity is that compliance to a desired behaviour can be measured and reported in a variety of different ways. In addition, interventions designed to target behaviour can be implemented at either an individual or group level leading to trials with varying layers of clustering.

To handle heterogeneous outcomes we can either convert all effect estimates to a common scale (e.g. using standardised mean differences) or have separate meta-analyses for different types of outcome measure (binary and continuous measures). To address the clustering structure, adjusted standard errors can be used with the inverse variance method, or weights can be assigned based on a consistent level of clustering, such as the number of healthcare professionals. A graphical method, the albatross plot utilises reported p-values only, and can synthesise data with both heterogeneous outcomes and clustering with minimal assumption and data manipulation.

Based on these methods, we reanalysed our data in four different ways and have discussed the strengths and weaknesses of each approach.

Keywords

Behaviour change, evidence synthesis, trials.

Background

In healthcare, many complex interventions are designed with the aim of changing the behaviour of individuals or groups of individuals. When designing new interventions, it is helpful to know which behaviour change techniques are most effective, and in which context. The behaviour change technique taxonomy¹ has identified and classified 93 different behaviour techniques, and each of these may be used alongside other techniques to form a complex intervention. The types of behaviours that these interventions could be targeting are numerous and varied; health behaviours such as eating a low calorie diet, or ceasing smoking; or clinical behaviours such as following government guidelines, prescribing drugs or washing hands. When summarising behaviour change research, one option would be to consider the effect of a specific intervention on a specific behaviour, but the large number of targeted behaviours would lead to a huge number of potential systematic reviews (or comparisons within a systematic review); each aiming to answer a different question, but unlikely to have enough statistical power to do so. In addition, it may be hard to interpret evidence from multiple systematic reviews of similar interventions that report conflicting conclusions, and present evidence in different ways. As described by Melendez-Torez², there are situations where it makes sense to group together interventions as ‘clinically meaningful units’ with a similar expected ‘theory of change’. In terms of behaviour change techniques, it can be informative to combine evidence to answer a broad question about how well a particular behaviour change technique (or group of techniques) has performed, on average, on any type of behaviour, and to use this information to identify which techniques are effective. This can be supplemented with analysis of effect moderators, to identify the contexts in which the technique is more or less effective.

Interventions to change healthcare professional (HCP) behaviour can be designed to target the individual HCP, or team of HCPs. Trials of this type of intervention can vary in terms of the unit of randomisation which can be either the individual HCP, or a group of HCPs such as those working within the same site (surgery, nursing home, ward, hospital). The unit of analysis in these trials can also vary and is not necessarily the same as the unit of randomisation; for example with randomisation at the level of GP surgery but with data recorded for each individual patient. The outcomes could be measured using a variety of denominators, such as the individual patient (e.g. binary measure of whether a test was ordered), individual HCPs (e.g. number

of tests ordered per GP), or at site-level (e.g. proportion of patients with an appropriate test order on a hospital ward). These multiple and varied layers need to be considered in terms of adjustment for clustering, combination of data and interpretation of results.

There are several proposed methods of summarising mixed measures of behavioural outcome. Higgins et al.³ provide an overview of methods to synthesise quantitative evidence in systematic reviews of complex health interventions. They describe and compare a number of graphical methods to combine different outcomes; as well as synthesis methods using effect size estimates, which are suitable for complex interventions. One approach is to combine effect sizes (standardised mean differences (SMDs)) using standard errors to derive weights for the studies in meta-analysis. In addition to allowing for different measurements (both binary and continuous) to be combined, this approach can accommodate a mixture of individually randomised and cluster randomised trials using weights based on adjusted standard errors. In some systematic reviews, binary measures of the same outcome are analysed and reported separately from continuous ones^{4,5}. An alternative approach⁶ to using weights based on standard errors is to use study weights based on the number of health professionals included in the study. The albatross plot⁷ is a graphical method which allows synthesis of summary data in a variety of formats, using only p-values plotted against sample size; this can be used to assess the consistency of results visually and allows estimation of average effect sizes.

Aims

Our aim was to examine methods for conducting meta-analysis in the context of heterogeneous behavioural outcome measures with clustering using a case study. We have applied different methods to data from the SOCIAL systematic review⁸⁻¹⁰ to illustrate the methods and examine their strengths and weaknesses.

Dataset

SOCIAL^{89,10} was a systematic review of randomised controlled trials, looking at the effect of social norms interventions on the clinical behaviour of health care workers, where a social norms intervention is defined as ‘an intervention which aims to change the behaviour of an individual by exposing them to the values, beliefs, attitudes or behaviours of a reference group or person’ (Tang, 2021, p.2) This review looked at the effects of any social norms intervention on any type of clinical behaviour, and aimed to answer an overall question about the effectiveness of social norms interventions, as well as more specific questions related to different types of intervention, settings, contexts and behaviour.

The SOCIAL systematic review included 102 unique trials that assessed the effect of a social norms intervention on the clinical behaviour of health workers. For ease of presentation, here we focus on a subset: 16 trials that assessed the effect of ‘credible source’ interventions either alone or alongside other interventions. A credible source intervention provides communication either in favour of or against a particular behaviour by a person generally agreed on as credible with the aim of persuading the recipient¹. For example Hallsworth¹¹ include a persuasive letter from the Chief Medical Officer in their intervention to reduce antibiotic prescriptions amongst high prescribing GPs.

Note that 2 of these 16 trials had more than two arms that tested the effect of a credible source intervention, so there were 18 different comparisons included. Table 1 shows the units of randomisation and analysis and how they vary by study.

The SOCIAL review found that social norms interventions appeared to be an effective method of changing the clinical behaviour of healthcare workers, with credible source interventions appearing to be most effective on average¹⁰.

Analysis

All analyses were performed using STATA 14.0¹². We have reported results of both fixed and random effects meta-analysis.

Where some summary data were missing for the reported trial results, we have imputed missing information (e.g. standard deviations or intra-cluster correlation coefficients (ICCs)) using other information in the trial

paper or values from similar trials¹³. Sensitivity to imputed values was assessed by imputing a range of different values.

Where the reported outcome data were from either an individually randomised trial or a cluster trial where the results had already been adjusted for clustering by the unit of randomisation, the standard errors were utilised without adjustment. Where adjustment for clustering was required the standard error was multiplied by the square root of the design effect (DE); this requires the average cluster size (M) and the intra-class correlation coefficient (ICC)¹³.

$$DE = 1+(M-1)ICC$$

Where possible we report the I² statistic as a measure of heterogeneity¹³. The I² statistic estimates the percentage of variation across studies that is due to study heterogeneity rather than chance.

Method 1: Standardised Mean Differences, weights based on adjusted standard error

This method is commonly used, including in the SOCIAL systematic review⁸ and other reviews¹⁴⁻¹⁷. This method is simple to use, it utilises information that is generally reported, and it can be performed using standard statistical software. All reported measures of intervention effect are converted into an approximation of the standardised mean difference (SMD) using the formulae in Table 2^{18,19}.

The formula for a standardised mean difference for a continuous outcome (Table 2) refers to Cohen’s d for ease of calculation. As an alternative Hedges g may be used¹⁹ which allows a correction for small sample size.

We applied these methods to the SOCIAL meta-analysis using the inverse of the squared adjusted standard error as weights (inverse variance method¹³).

Where a trial reported both a continuous and binary outcome measure with appropriately adjusted standard errors, we utilised the continuous measure but also calculated the SMDs and standard errors using the binary measure to check for anomalies. Note that rules such as this should be pre-specified to avoid post-hoc decisions that could introduce bias.

Method 2: Separate analyses for binary and continuous outcomes, weights based on adjusted standard errors

In this method two separate analyses are produced for the same outcome; one for those that were reported as binary measurements and one for those that were reported as continuous measurements. This method has been used in a number of systematic reviews of health behaviour change^{4,5,16}. This method requires very little data manipulation and adopts a conservative approach to heterogeneity by keeping the two types of outcome separate.

For illustration we performed meta-analysis on the SOCIAL data using odds ratios for binary data and standardised mean differences for continuous data, using the inverse variance method with weights based on adjusted standard errors.

Outcomes were reported as continuous measures on a variety of different scales; therefore they were converted to standardised mean differences as above²⁰. If all continuous measures had been measured on the same scale, e.g. mean percentage on a scale of 0 to 100, they could be meta-analysed using means and standard deviations. For meta-analysis of binary data, all summaries need to be converted to the same format (odds ratio, risk ratio or risk difference); it is recommended that this be chosen in advance at the protocol stage to avoid selective reporting. We chose odds ratios¹⁸ here as they have certain desirable mathematical properties; their symmetrical nature would mean that an analysis where the outcome measure is ‘compliance’ or an analysis where the outcome measure is ‘non-compliance’ would lead to identical conclusions.

Sometimes trials report the same outcome measure in both binary and continuous formats – for example in a trial where the desired behaviour is ‘test ordering’; summary data could be reported both in terms of the overall proportion of patients who had a test ordered, and the mean proportion of tests ordered by health

care professional. Where a trial has reported an outcome in both binary and continuous formats, we included both measures in the two separate meta-analyses. Note that when using this method, continuous and binary results may not be later combined together as this would lead to double counting of the same participants.

Method 3: SMDs, weighting by number of health care professionals

Where the population of interest is the health care professional, it may be desirable to weight the results by the number of health care professionals included^{6,21} to aid population inference. This method utilises commonly reported summary information without adjustment for clustering. In this method of meta-analysis the studies are weighted by the number of health care professionals as an alternative to the commonly used inverse variance method which uses weights based on standard errors. As pointed out in table 4 there are weaknesses to this approach which we discuss below.

Not every trial in the SOCIAL review reported the number of health care professionals – for example a cluster trial where an intervention was directed at all staff on a hospital ward. Where no information was given about the number of health care professionals, Ivers et al.⁶ used the number of practices/hospitals/communities instead, and we followed that method here. An alternative might be to estimate the number of health care professionals using data from similar studies – e.g. using mean number of GPs per surgery or mean number of nursing staff on a hospital ward. Note that this method needs to be combined with method 1 or method 2 above or an alternative way of summarising mixed outcome measures; here we combined it with method 1 to summarise standardised mean differences.

Method 4: Albatross plots

The albatross plot was first described by Harrison et al.⁷ and is also discussed in Higgins²². This method requires minimal data extraction or manipulation and allows data to be synthesised even in circumstances when outcomes are reported in multiple different formats or where no summary statistics are reported. Reported results are split into two groups according to the direction of effect; and then p-values are plotted against sample size. Where necessary, 1-sided p-values need to be converted to 2-sided p-values (or vice versa) to ensure consistency. An albatross plot allows us to combine outcome data that was reported in a variety of different ways, including from studies where only a p-value was provided. Under an assumption of normality, you would expect results corresponding to the same effect size to lie along a contour, with p-values generally getting smaller as sample size increases. Contours can be added to the plot for a range of different effect sizes based on standardised mean differences, mean differences, odds ratios or other summary of choice. Effect sizes can be estimated according to where the majority of points lie. We have added contours to represent standardised mean differences of 0.3, 0.6 and 0.9. Heterogeneity can also be explored visually by looking at how closely trials tend to group together along a particular contour.

Note that where p-values are obtained from studies that are clustered in some way, adjustment of sample size is necessary. One method of doing this is to calculate the effective sample size (E) using the sample size (S), the reported intra-class correlation coefficient (ICC) and the average cluster size (M) using the formula²³

$$E = \frac{S}{1 + ICC \times (M - 1)}$$

An alternative is to replace the sample size with the number of health care professionals (or sites) as in method 3.

For illustration we produced a contour plot using the number of health care professionals (or sites) as the sample size (Figure 1)

Results

Method 1 produced pooled SMDs of 0.14 (95% CI 0.10 to 0.17) and 0.31 (95% CI 0.14 to 0.51) for the fixed and random effects results. There is a marked difference between the results for fixed and random effects;

with the fixed result having a smaller effect size and tighter confidence interval; this is because the fixed effects analysis gives more weight to large trials, which tended to have more modest effect sizes (Table 3).

Method 2 resulted in pooled OR 1.13 (95% CI 1.06 to 1.20) and pooled SMD 0.50 (95% CI 0.42 to 0.59) for fixed effects; OR 1.13 (95% CI 1.06 to 1.20) and SMD 0.92 (95% CI 0.11 to 1.73) for random effects. One study contributed data to both the odds ratio and SMD estimate. The method using odds ratios produced a far less heterogeneous result than that for the SMDs in this case but as they are from different sets of trials it is difficult to infer why.

Method 3 resulted in an SMD 0.57 (95% CI 0.50 to 0.64). This weighted average produced the narrowest confidence intervals for SMDs.

For Method 4, we can see from Figure 1 that all studies reported a positive effect so it is clear that, on average, credible source interventions seem effective. The fact that the points are not clustered around one particular contour line tells us that there is a high level of heterogeneity. Both large and small studies appear to be associated with very small p-values and large effect sizes, so there is little evidence of publication bias.

Three of the methods produced an SMD, which ranged from 0.14 to 0.57. All were statistically significant, suggesting that we can be reasonably confident that a positive effect exists, but less confident in estimating the size of the effect as it is sensitive to the method chosen.

Challenges

In Table 4 we summarise the strengths and weaknesses of the different approaches. In systematic reviews of complex interventions there is likely to be a large amount of heterogeneity due to differences in setting, population, intervention and study design. When combining different types of outcomes, measured and reported in a variety of different ways, heterogeneity due to outcome measurement also has to be a serious additional consideration. Our estimate of heterogeneity, I^2 for the SMD analyses ranged from 95.3% to 98.5% suggesting substantial heterogeneity. Exploration of heterogeneity is not the focus of the paper and has been discussed by a number of authors^{24,25,26}. Sources of heterogeneity can be explored using methods such as subgroup analysis²⁷ and meta-regression²⁸ although these common approaches are subject to ecological fallacy, and superior approaches exist where sufficient data are available²⁹. In contrast to a meta-analysis of a well-defined pharmaceutical intervention, where heterogeneity is generally seen as a nuisance, identifying the sources of the heterogeneity is often a key research questions when synthesising data from complex interventions.

Some authors have expressed concerns about the use of SMDs in meta-analysis. The SMD estimates the average improvement in outcome per SD on whatever scale that outcome is measured on; as Greenland³⁰ points out the SD measured within a trial is likely to be different to the population SD and will vary according to the design features of the trial (e.g. inclusion/exclusion criteria). Trials are often designed to minimise variability and therefore SDs reported are likely to be smaller than the SD in the target population, leading to an overestimate of the treatment effect of interest. Another problem, as discussed by Senn³¹, that is especially pertinent here, is that the SD will depend on the measurement error, and since we have lots of different measurement scales we will have lots of different measurement errors; this means that you could get lots of different SMDs even if the treatment effect was the same in each study.

In an attempt to combine all available information we have converted odds ratios into SMDs using the methods described by Chinn³². This method provides an estimate of the SMD from an odds ratio using the assumption that the odds ratio has come from a dichotomy of a normally distributed continuous variable; this may be a poor estimate when this assumption is not true. Sanchez-Meca³³ compares alternative indices to combine continuous measures with dichotomies and show that this method slightly underestimates the SMD. Our conclusions were unchanged when binary and continuous data were analysed separately, but the SMDs estimated from continuous data alone were considerably higher than those when binary data were combined so it is possible that by converting odds ratios to SMDs we were underestimating the true treatment effect in this context.

Varying units of randomisation and analysis lead to difficulties both in terms of synthesis methods and interpretation. One of our reported methods (Method 3) aims to apply consistent weighting based on the number of health care professionals to allow inference about a consistent population; however this leads to other problems. Weights based on sample size do not take into account the variability of the data, essentially assuming a constant standard deviation across all trials. In their simulation study, Marin-Martinez and Sanchez-Meca³⁴ show that weighting by the inverse variance yields less biased results than weighting by sample size. Complexity is added when a review wishes to combine evidence from different types of trial design^{35,36}. Individually randomised trials, cluster randomised trials and stepped wedge trials are all useful in answering questions about behaviour change interventions targeted at health care professionals, but you would not necessarily expect the SMD (effect size) to be consistent across each type of trial due to the different units of analysis (and therefore different underlying SDs)^{37,38}. Some consensus among trialists of health professional behaviour change interventions, in the form of a core outcome set³⁹ would be useful for future systematic reviews. Consistency in terms of outcomes used, unit of analysis and format of outcome reporting is desirable. In addition, we may want to separate out the effect on the health care provider from the effect on the individual patient; this would require individual participant data and multilevel modelling²⁴.

Some trials used in this analysis have reported ‘mean percentage compliance’ or similar – e.g. the percentage of occasions a test was ordered, averaged over a group of GPs. This measurement is bounded between 0% and 100% and therefore cannot be considered truly continuous. Inference methods (meta-analysis of SMDs) used here assume continuity and normality and are likely to perform poorly where results are close to the boundaries (0% and 100%). We performed additional sensitivity analyses removing trials where the mean compliance was between 0% and 20% or between 80% and 100%; and results appeared robust. Alternative methods to analyse proportions include those suggested by Miller⁴⁰ and Stijnen et al.⁴¹ and these may be preferable when meta-analysing proportions alone.

We acknowledge all of these challenges and feel that conclusions based on any of the methods presented here need to be very cautious. However we feel that there are occasions where the combination of mixed outcomes is still warranted, but should be accompanied with appropriate sensitivity analyses and caveats.

Conclusions

Systematic reviews of complex behaviour change interventions in healthcare may include a heterogeneous set of studies in terms of content, context, trial design and setting. The measures of behaviour change may also vary which leads to difficulty in attempts to synthesise the data, as well as increased heterogeneity.

In this paper we have presented 4 different methods for combining behavioural outcome measures from trials, described the strengths and weaknesses of each method, and the problems inherent with combining heterogeneous outcome measures with mixed levels of clustering. Each of the methods presented has advantages and disadvantages, summarised in table 4, and we recommend that reviewers chose their methods carefully based on the needs of their review, and plan methods and data conversion policies in advance to avoid selective reporting. We observed that for our data, conclusions would remain robust regardless of the methods of analysis chosen; however the estimated magnitude of the treatment effect varied quite markedly according to the method chosen. We view the methods presented as useful when trying to convert all outcome measures to the same scale and to provide an overall summary, but results should be interpreted extremely cautiously given the limitations. We would recommend that results are used as an aid in summarising the evidence and generating future hypotheses rather than to infer future effects.

Funding

This project is funded by the National Institute for Health Research (NIHR) Health Services and Delivery Research, reference 17/06/06 - The impact of social norms interventions on clinical behaviour change among healthcare workers: a systematic review. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflicts of interest/Competing interests

The authors have no conflicts of interest to declare that are relevant to the content of this article

Ethics approval (include appropriate approvals or waivers)

This article utilises only summary data from published trials. It is exempt from the need for ethical approval.

Availability of data and materials (data transparency)

Data to be stored on Figshare (DOI will be made available)

Code availability (software application or custom code)

Code to be stored on Figshare (DOI will be made available)

Authors' contributions

SR conceived the idea for the article, performed analyses and wrote the draft manuscript. SC led the original systematic review, developed the idea for the article and wrote sections of the article. SD and JW provided methodological input and substantially edited the article.

References

1. Michie S, Richardson M, Johnston M, et al. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine* . 2013;46(1):81-95. doi:10.1007/s12160-013-9486-6
2. Melendez-Torres GJ, Bonell C, Thomas J. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Med Res Methodol* . 2015;15:47-47. doi:10.1186/s12874-015-0040-z
3. Higgins JPT, López-López JA, Becker BJ, et al. Synthesising quantitative evidence in systematic reviews of complex health interventions. *BMJ Global Health* . 2019;4(Suppl 1):e000858. doi:10.1136/bmjgh-2018-000858
4. Davey P, Marwick CA, Scott CL, et al. Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database of Systematic Reviews* . 2017;(2)doi:10.1002/14651858.CD003543.pub4
5. Vaona A, Banzi R, Kwag KH, et al. E-learning for health professionals. *Cochrane Database of Systematic Reviews* . 2018;(1)doi:10.1002/14651858.CD011736.pub2
6. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews* . 2012;(6)doi:10.1002/14651858.CD000259.pub3
7. Harrison S, Jones HE, Martin RM, Lewis SJ, Higgins JPT. The albatross plot: A novel graphical tool for presenting results of diversely reported studies in a systematic review. *Research Synthesis Methods* . 2017;8(3):281-289. doi:10.1002/jrsm.1239
8. Cotterill S, Powell R, Rhodes S, et al. The impact of social norms interventions on clinical behaviour change among health workers: protocol for a systematic review and meta-analysis. *Systematic Reviews* . 2019/07/18 2019;8(1):176. doi:10.1186/s13643-019-1077-6
9. Cotterill S, Tang MY, Powell R, et al. Social norms interventions to change clinical behaviour in health workers: a systematic review and meta-analysis. 2020;8:41. doi:10.3310/hsdr08410
10. Tang MY, Rhodes S, Powell R, et al. How effective are social norms interventions in changing the clinical behaviours of healthcare workers? A systematic review and meta-analysis. *Implementation Science* . 2021/01/07 2021;16(1):8. doi:10.1186/s13012-020-01072-1
11. Hallsworth M, Chadborn T, Sallis A, et al. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. *Lancet* . Apr 23 2016;387(10029):1743-52. doi:10.1016/S0140-6736(16)00215-4
12. Stata C. *Stata release 14* . 2015.

13. Higgins JPT, Cochrane C. *Cochrane handbook for systematic reviews of interventions* . 2019.
14. Murray JM, Brennan SF, French DP, Patterson CC, Kee F, Hunter RF. Effectiveness of physical activity interventions in achieving behaviour change maintenance in young and middle aged adults: A systematic review and meta-analysis. *Soc Sci Med* . Nov 2017;192:125-133. doi:10.1016/j.socscimed.2017.09.021
15. Grimmer C, Corbett T, Brunet J, et al. Systematic review and meta-analysis of maintenance of physical activity behaviour change in cancer survivors. *Int J Behav Nutr Phys Act* . Apr 27 2019;16(1):37. doi:10.1186/s12966-019-0787-4
16. Corepal R, Tully MA, Kee F, Miller SJ, Hunter RF. Behavioural incentive interventions for health behaviour change in young people (5-18years old): A systematic review and meta-analysis. *Prev Med* . May 2018;110:55-66. doi:10.1016/j.ypmed.2018.02.004
17. Baskerville NB, Liddy C, Hogg W. Systematic review and meta-analysis of practice facilitation within primary care settings. *Annals of family medicine* . Jan-Feb 2012;10(1):63-74. doi:10.1370/afm.1312
18. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ (Clinical research ed)* . 2000;320(7247):1468-1468. doi:10.1136/bmj.320.7247.1468
19. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ* . 2019;364:k4817. doi:10.1136/bmj.k4817
20. Saramago P, Woods B, Weatherly H, et al. Methods for network meta-analysis of continuous outcomes using individual patient data: a case study in acupuncture for chronic pain. *Bmc Med Res Methodol* . 2016/10/06 2016;16(1):131. doi:10.1186/s12874-016-0224-1
21. Tuti T, Nzinga J, Njoroge M, et al. A systematic review of electronic audit and feedback: intervention effectiveness and use of behaviour change theory. *Implementation Science* . 2017/05/12 2017;12(1):61. doi:10.1186/s13012-017-0590-z
22. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Stat Med* . Sep 30 2013;32(22):3926-3943. doi:10.1002/sim.5831
23. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* . Jun 1992;48(2):577-85.
24. Petticrew M, Rehfuess E, Noyes J, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *Journal of Clinical Epidemiology* . 2013/11/01/ 2013;66(11):1230-1243. doi:https://doi.org/10.1016/j.jclinepi.2013.06.005
25. Davis J, Mengersen K, Bennett S, Mazerolle L. Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus* . 2014/09/10 2014;3(1):511. doi:10.1186/2193-1801-3-511
26. Tanner-Smith EE, Grant S. Meta-Analysis of Complex Interventions. *Annual Review of Public Health* . 2018;39(1):135-151. doi:10.1146/annurev-publhealth-040617-014112
27. Borenstein M, Higgins JP. Meta-analysis and subgroups. *Prevention science : the official journal of the Society for Prevention Research* . Apr 2013;14(2):134-43. doi:10.1007/s11121-013-0377-7
28. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine* . Jun 15 2002;21(11):1559-73. doi:10.1002/sim.1187
29. Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* . 2017;356:j573. doi:10.1136/bmj.j573
30. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology* . Sep 1991;2(5):387-92.

31. Senn S. U is for unease: reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research* . 2011;3(2):302-309. doi:10.1198/sbr.2010.10024

32. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* . Nov 30 2000;19(22):3127-31. doi:10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.0.co;2-m

33. Sanchez-Meca J, Marin-Martinez F, Chacon-Moscoso S. Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological methods* . Dec 2003;8(4):448-67. doi:10.1037/1082-989x.8.4.448

34. Marin-Martinez F, Sanchez-Meca J. Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. *Educational and Psychological Measurement* . 2010/02/01 2009;70(1):56-73. doi:10.1177/0013164409344534

35. Laopaiboon M. Meta-analyses involving cluster randomization trials: a review of published literature in health care. *Stat Methods Med Res* . 2003;12(6):515-530. doi:10.1191/0962280203sm347oa

36. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Stat Med* . 2002;21(19):2971-2980. doi:10.1002/sim.1301

37. Walwyn R, Roberts C. Meta-analysis of standardised mean differences from randomised trials with treatment-related clustering associated with care providers. *Stat Med* . Mar 30 2017;36(7):1043-1067. doi:10.1002/sim.7186

38. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Statistical methods in medical research* . 2001/10/01 2001;10(5):325-338. doi:10.1177/096228020101000502

39. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. *Trials* . 2017/06/20 2017;18(3):280. doi:10.1186/s13063-017-1978-4

40. Miller JJ. The Inverse of the Freeman – Tukey Double Arcsine Transformation. *The American Statistician* . 1978/11/01 1978;32(4):138-138. doi:10.1080/00031305.1978.10479283

41. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med* . 2010;29(29):3046-3067. doi:10.1002/sim.4040

Table 1: Units of randomisation and analysis for the 18 credible source comparison

	Number of studies	Number
Unit of randomisation Patient Health care professional Site (ward, hospital, surgery etc)	0 2 14	0 2 16
Unit of analysis Patient Health care professional Site (ward, hospital, surgery etc)	8 4 4	10 4 4

Table 2: Formulae to convert extracted data to SMDs

Type of outcome measure	Data to extract	Standardised mean difference (d)	Standard error of standardised mean difference
Continuous reported as mean or mean difference	Means (M_1 and M_2), standard deviations (S_1 and S_2) and sample size per group (n_1 and n_2)	$\frac{M_1 - M_2}{S}$ <p>where $S = \sqrt{\frac{(n_1 - 1)^2 S_1 + (n_2 - 1)^2 S_2}{n_1 + n_2 - 2}}$</p>	$\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{n_1 + n_2 - 2}}$

Type of outcome measure	Data to extract	Standardised mean difference (d)	Standard error of standardised mean difference
Binary reported using odds ratios	Natural logarithm of odds ratio (lnOR) and standard error of log odds ratio $SE_{\ln OR}$. This can be obtained from a 95% confidence interval for the odds ratio by taking natural logs and dividing by 2×1.96	$\frac{\sqrt{3}}{\pi} \ln OR$	$\frac{\sqrt{3}}{\pi} SE_{\ln OR}$
Raw binary data	Raw binary data (c_1/n_1 and c_2/n_2) where c_1 and c_2 are the number of participants complying with the behaviour of interest by group.	$OR = \frac{\frac{c_1}{n_1 - c_1}}{\frac{c_2}{n_2 - c_2}}$ Take natural log and continue as above	$SE_{\ln OR} = \sqrt{\frac{1}{c_1} + \frac{1}{n_1 - c_1} + \frac{1}{c_2} + \frac{1}{n_2 - c_2}}$ Continue as above

Table 3: Summary of results for credible source data by 5 different methods

Method	Number of comparisons	Result	Measure of heterogeneity
Method 1 SMDs. Weights based on adjusted standard errors	18	Fixed effects SMD 0.14(95% CI 0.10 to 0.17) Random effects SMD 0.31(95% CI 0.14 to 0.51)	$I^2 = 95.3\%$
Method 2 Separate analyses for binary and continuous data. Weights based on adjusted standard errors	OR 12 SMD 7	Fixed effects OR 1.13(95% CI 1.06 to 1.20) SMD 0.50(95% CI 0.42 to 0.59) Random effects OR 1.13(95% CI 1.06 to 1.20) SMD 0.92(95% CI 0.11 to 1.73)	$I^2 = 0\%$ $I^2 = 98.0\%$
Method 3 SMSs. Weighting by number of HCP	18	SMD 0.57(0.50 to 0.64)	$I^2 = 98.5\%$
Method 4 Albatross plot	18	All studies reported a positive effect so clear evidence of treatment effect.	Points not clustered around a single contour line so high levels of heterogeneity

Table 4 Strengths and weaknesses of each approach

Approach	Strengths	Weaknesses
Method 1 SMDs. Weights based on adjusted standard errors	All available data combined Clustering accounted for at level of randomisation	Mixture of different outcomes and formats likely to lead to heterogeneity May be difficult to interpret Inconsistent units of analysis (patient/HCP/site) Estimation assumptions may not hold
Method 2 Separate analyses for binary and continuous data. Weights based on adjusted standard errors	Likely to lead to less heterogeneity than method 1 as more similar measures are being combined. Little manipulation or estimation required	Does not combine all available information in a single analysis, which leads to loss of power and multiplicity Two analyses may give conflicting results
Method 3 SMDs. Weighting by number of HCP	Consistent units – weighted by health care professional	Number of health care professionals not always reported, requiring an estimate to be imputed Weighting may be related to quality of reporting; e.g. poorly reported studies get less weight. Unit of analysis error when not randomised at level of analysis Weights related to the size of the study but not the variability/precision Issues with SMDs as above
Method 4 Albatross plot	May include additional studies that report p-value only No assumptions	Difficult to check that p-values are correct if not accompanied by other summary data P-values prone to selective reporting Need to adjust sample size in some way for cluster trials

Figure Albatross plot using ‘number of health care professionals’ as sample size.

