Essay 1 Integrating multiple sources of evidence: a Bayesian perspective

Samuel I Watson and Richard J Lilford

Warwick Medical School, University of Warwick, Coventry, UK

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

Published May 2016
DOI: 10.3310/hsdr04160-1

This essay should be referenced as follows:

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

List of figures

List of abbreviations

CPOE	computerised physician order entry
RCT	randomised controlled trial

Abstract

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

Scientific summary

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

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

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

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

The purpose of evidence synthesis

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

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

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

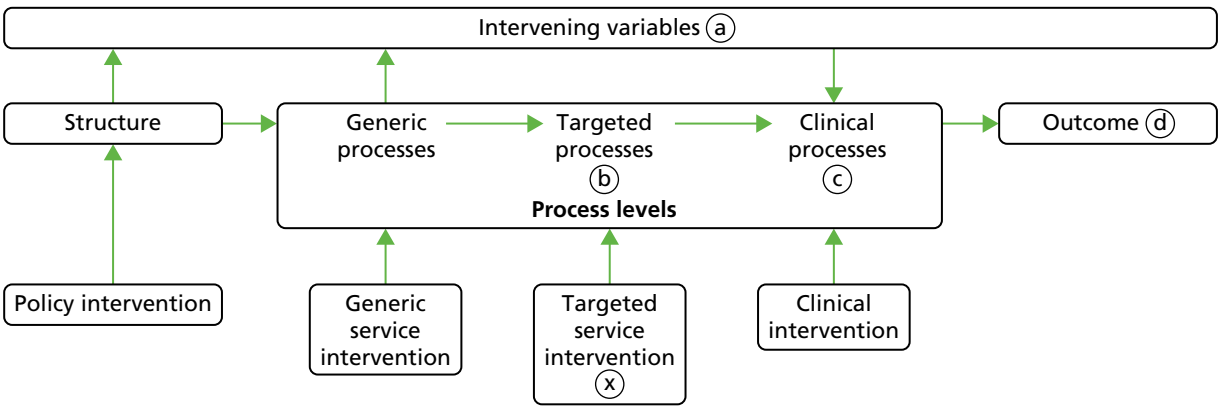

FIGURE 1.1 Causal chain showing how interventions at different levels may impact on downstream processes and outcomes. Reproduced from Evaluating policy and service interventions: framework to guide selection and interpretation of study end points, Lilford RJ, Chilton PJ, Hemming K, Girling AJ, Taylor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

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

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

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

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

The nature of the system

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

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

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

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

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

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

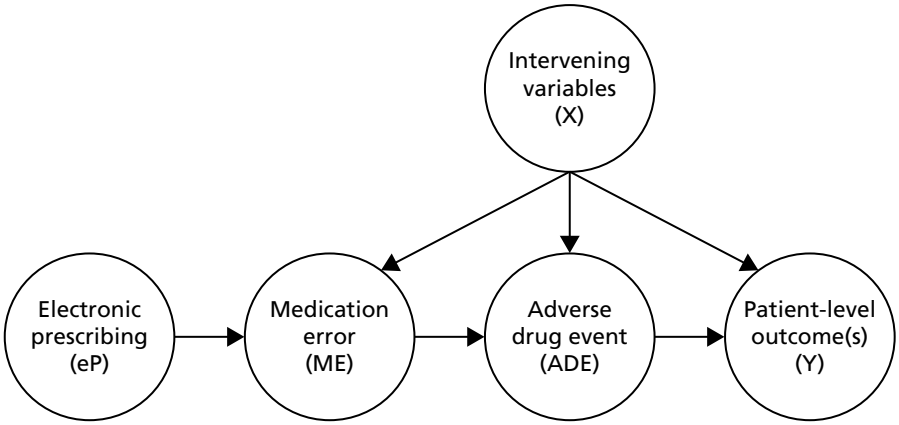

FIGURE 1.2 Qualitative causal model showing the effects of an electronic prescribing system on patient-level outcomes.

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

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

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

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

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

resembling childbed fever after puncturing a finger during an autopsy, Semmelweis inferred that 'cadaveric matter' was the cause. To test this he had medical students disinfect their hands after performing autopsies and the mortality rate dropped significantly.

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

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

Methods of evidence synthesis

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

Theory

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

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

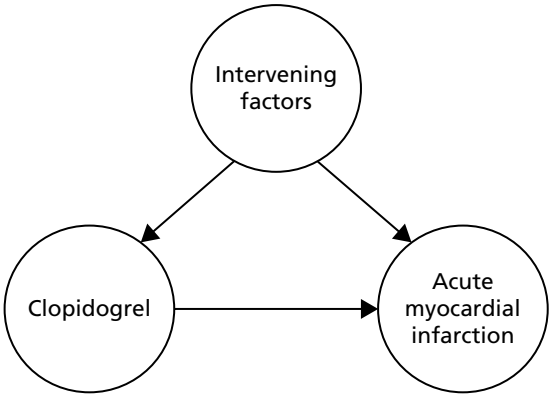

FIGURE 1.3 Qualitative causal model for the effects of clopidogrel on acute myocardial infarction.

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

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

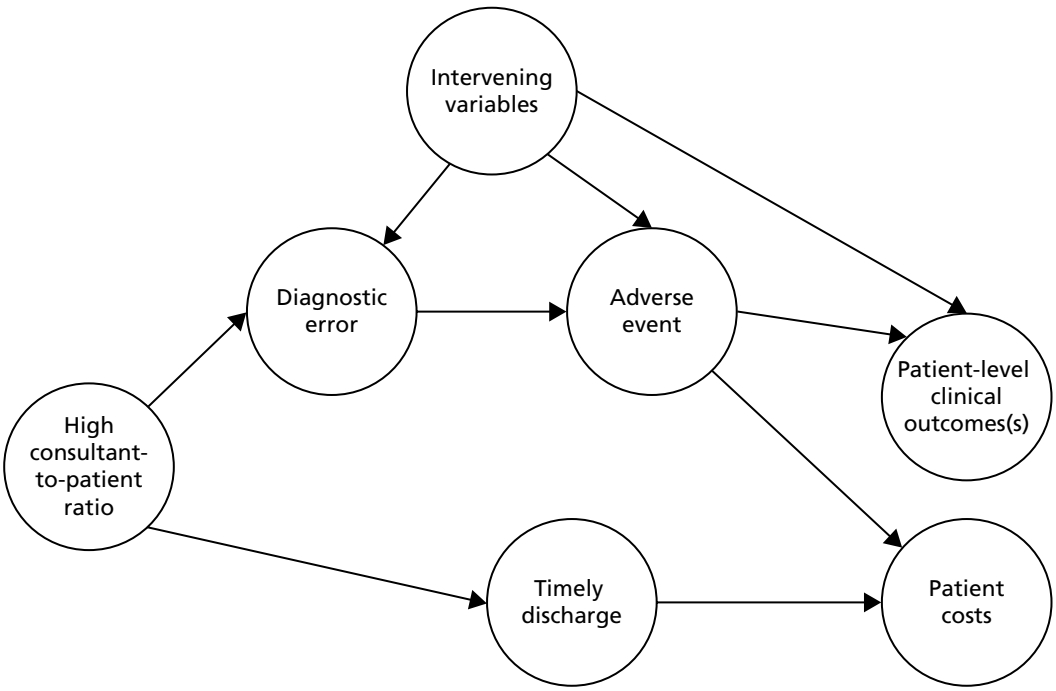

FIGURE 1.4 Qualitative causal model.

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

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

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

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

Identifying and evaluating evidence in the literature

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

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

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

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

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

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

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

study design, heterogeneity, or inclusion of studies of pharmacological interventions'.[36] That said, in some cases there are considerable differences between the results of randomised and non-randomised studies of the same intervention.[37,38]

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

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

Evidence synthesis

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

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

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

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

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

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

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

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

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

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

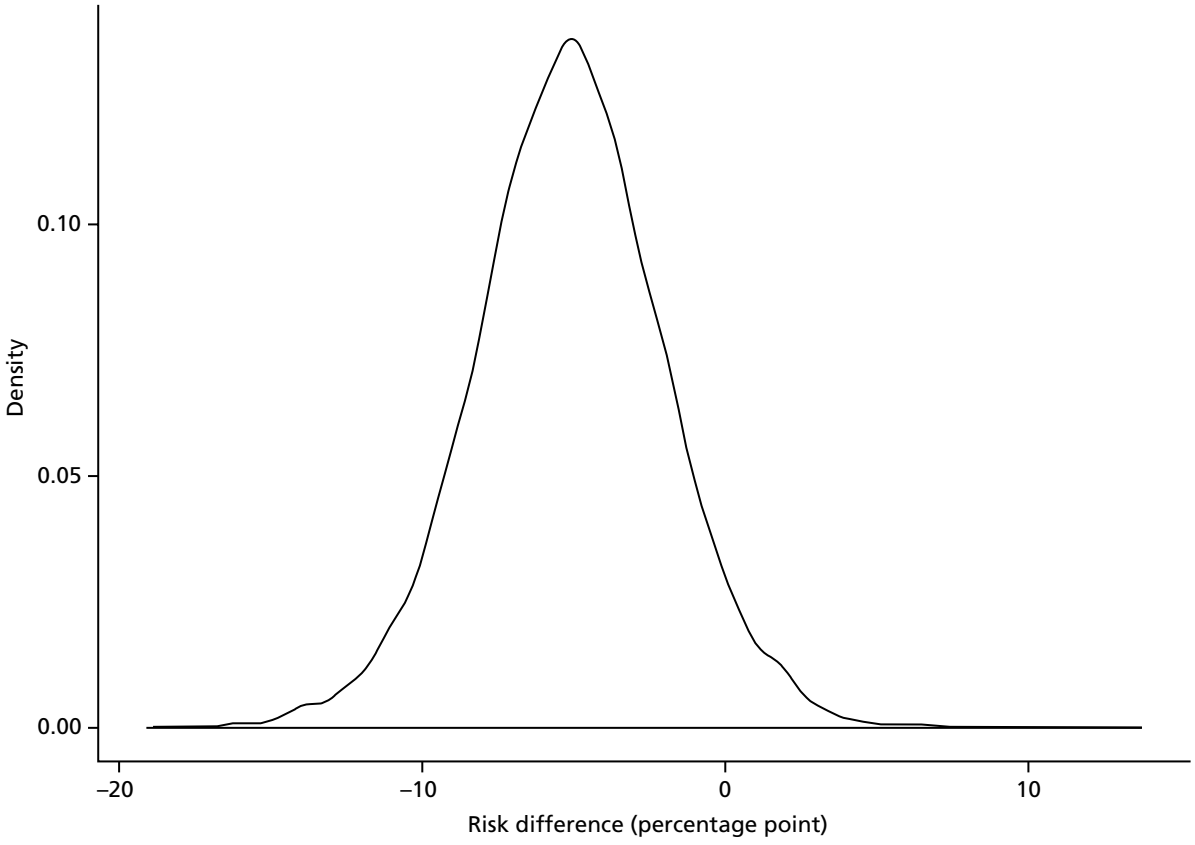

FIGURE 1.5 Posterior distribution from a random-effects meta-analysis of the absolute effect of CPOE on the risk of a patient experiencing a medication error compared with prescribing with a paper-based system.

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

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

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

Decision analysis

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

Conclusions

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

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

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

Acknowledgements

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

Contributions of authors

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

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

Both authors approved the final version of the manuscript.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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