Synthesizing Evidence of Risk

Edoardo Aromataris

Lisa Hopp

Zachary Munn

Synthesizing Evidence of Risk

Edoardo Aromataris

Lisa Hopp

Zachary Munn

DISCLAIMER

Care has been taken to confirm the accuracy of the information present and to describe generally accepted practices. However, the authors, editors, and publisher are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, expressed or implied, with respect to the currency, completeness, or accuracy of the contents of the publication. Application of this information in a particular situation remains the professional responsibility of the practitioner.

To purchase additional copies of this book, please visit Lippincott's NursingCenter.com or call our customer service department at (800) 638-3030 or fax orders to (301) 223-2320. International customers should call (301) 223-2300.

Visit Lippincott Williams & Wilkins on the Internet: http://www.lww.com. Lippincott Williams & Wilkins customer service representatives are available from 8:30 am to 6:00 pm, EST.

Series Editor: Professor Alan Pearson AM

This series of concise texts is designed to provide a "toolkit" on synthesizing evidence for healthcare decision-making and for translating evidence in action in both policy and practice. The series seeks to expand understandings of the basis of evidence-based health care and brings together an international range of contributors to describe, discuss and debate critical issues in the field.

Incredible developments have occurred in the synthesis and use of evidence in health care over the last several years, but the science and emerging practices that underpin evidence based health care are often poorly understood by policy makers and health professionals. Several emerging and exciting developments have much to offer health professionals. Firstly, new, deeper understandings of the nature of evidence and of ways to appraise and synthesise it have led to the development of more sophisticated methodologies for synthesis science. Secondly, the realisation that the rapid increase in the availability of high quality evidence has not been matched by increases in the translation of this evidence into policy and/or clinical action has spurred on developments in the science of knowledge implementation and practice improvement.

The burgeoning publications in this area – particularly books on evidence based health care - can go only so far in informing responsible and conscientious policy makers and health care practitioners. This new series Lippincott/Joanna Briggs Institute, "Synthesis Science in Health care", is devoted to communicating these exciting new interventions to researchers, clinicians on the frontline of practice and policy makers.

The books in this series contain step-by-step detailed discussions and practical processes for assessing, pooling, disseminating and using the best available international evidence. In all health care systems, the growing consensus is that evidence-based practice offers the most responsible course of action for improving health outcomes. All clinicians and health scientists want to provide the best possible care for patients, families and communities. In this series, our aim is to close the evidence to action gap and make that possible.

About the Authors

Dr Edoardo Aromataris B.Sc.(Hons), PhD is Associate Director for Synthesis Science in the Joanna Briggs Institute, Faculty of Health Sciences, the University of Adelaide. He is a bioscientist whose wish to see his research efforts have a direct impact on patient wellbeing led him from the laboratory bench to the field of evidence-based health care.

Dr Aromataris frequently lectures on methods of evidence synthesis both in Australia and abroad. He was appointed as Editor of the Joanna Briggs Institute's Library of Systematic Reviews in 2011.

Dr Lisa Hopp PhD, RN is Director of the Indiana Center for Evidence Based Nursing Practice: A Collaborating Centre of the Joanna Briggs institute. She is also professor of nursing at Purdue University Calumet where the Center is housed. She conducts systematic review training and conducts and publishes systematic reviews.

Her passion is to contribute to moving evidence into practice through evidence synthesis, educating clinical nurse specialists and collaborating with clinicians. She lectures in the United States and abroad on issues related to evidence based practice.

Zachary Munn BMedRad, GradDip HlthSc is a Research Fellow in Translation Science in the Joanna Briggs Institute, Faculty of Health Sciences, the University of Adelaide. He has a background in medical radiations and his particular interests focus on the translation of evidence into practice and in the systems and processes that can be utilised to achieve this. He has a keen interest in varying research methodologies, and is involved in a range of research and teaching activities within the Joanna Briggs Institute.

Contents

Introduction

I n the fields of health and social care, the synthesis of the best available evidence to support decision making at the policy and practice levels is increasing in importance. In health care, practitioners and patients make numerous decisions and, in doing so, weigh up numerous types of information before taking action (including the results of well-designed research; information related to patients/clients and their relevant others; the practitioner's own experiences; and the nature and norms of the setting and culture in which the care is being delivered).

Some serious consideration has been given to the meaning of evidence in relation to evidence-based health and social care and there is a growing literature on the evidence that can be incorporated and relied upon to inform practice. The term evidence refers to the basis of belief; the substantiation or confirmation that is needed in order to believe that something is true. Health professionals seek evidence to substantiate the worth of a very wide range of activities and interventions and thus the type of evidence needed depends on the nature of the activity and its purpose. Many of the methods that are applied to the broad range of evidence to inform health care practice and policy today, have their roots in evidence based medicine and investigating quantitative evidence to infer a causal relationship between some treatment or therapy and measurable outcome.

Synthesis of evidence related to risk is the primary focus of this volume. In the context of informing decisions in health care, risk refers to a deleterious or adverse outcome, primarily from 'everyday' or inadvertent exposure to a variable or risk factor. In some cases the 'risk' may arise from a treatment administered during the course of patient care also. Synthesis of this type of evidence is increasing in importance and relevance; evidenced by the increasing number of systematic reviews and meta-analyses appearing in publication. Information about risks, even if they appear only minimal, are of great interest to the general public and can have a marked impact on public health (Blettner et al., 1999). Research, investigation and critical analysis across this important field of inquiry has the potential to better inform all of us, health care professionals, policy makers or lay persons, about the decisions we make in our day to day lives and the impact those decisions may have on our health both now and in our future.

1

The Synthesis of Evidence of Risk

Chapter 1:
The Nature of Risk

Risk is a term we are all familiar with and most of us would agree that there is an element of 'risk' in almost everything we do as we go about our lives. All the activities we undertake in our day-to-day existence expose us to risk of varying degrees. Furthermore, all of us, often inadvertently, quantify risk and assign our own 'level' of risk to what it is we are doing. An activity for one individual, such as crossing a busy road where there is no pedestrian crossing, may be acknowledged as 'less risky' by that same individual if not attempted in peak hour traffic. Conversely, another individual may view the undertaking of crossing the road as not risky at all, perhaps due to characteristics of personality or past experience in activities they, and indeed most of us, would view as truly risky, for example base jumping or motor sport.

Despite this ubiquitous nature of risk, there is surprisingly little consensus as to how individuals understand and define the term and its use socially varies markedly as do nuances in its meaning. This variation of the understanding of the term and its use in our vocabulary and syntax, for the most part, appears aligned to the profession of the person using the term. For example, risk and its mitigation underlies much of the daily fervor of activity in the financial markets, with those involved assessing the tradeoff between the potential for large gains against the risk of potentially larger losses of a particular investment strategy; at the same time, a family physician may be explaining to their patient the increased risk of developing lung cancer their smoking habit is predisposing them to. These examples help illustrate the most common meaning of the term risk which, irrespective of the profession or experience of the person, be they occupied in finance, health care or some other pursuit, connotes the chance or probability of some hazardous or deleterious outcome as a consequence of the course of action in question.

Information regarding the risk or probability that a particular exposure causes or results in a disease or health problem is invaluable information for health care professionals and also policy makers. Synthesis of evidence related to etiology, or cause of disease, is the primary focus of this volume. Health care professionals use such evidence in their day to day decision making to provide informed advice to patients about their own self management of their health and well being and also, importantly, to assess and inform the likelihood that some previous exposure could help explain the cause of a presenting patient's current health predicament. Similarly,

government policy that claims to be informed by evidence must consider such information across a range of health and social policy decisions.

It is worth noting from the outset, in the context of health care and how we evaluate quantitative data using statistical methods, the term risk takes on a dual meaning. Health care professionals use risk to indicate the likelihood or probability of a beneficial outcome occurring as well as a bad. For example, the risk of mortality due to cardiovascular events in hypertensive individuals receiving drug therapy with appropriate medication can be quantified. When it is compared to a similar population who are not treated, the risk (or relative risk) of mortality will (hopefully!) appear reduced. In this example, this reduction in risk observed with the pharmacological intervention is most commonly established by experimental studies, preferably randomized controlled trials.

When considering quantitative evidence to inform health care (the distinction is made here in recognition of other volumes in this series addressing the synthesis of qualitative research, specifically Volume 2), the randomized controlled trial (RCT) is considered the pinnacle of evidence for an individual piece of research endeavor, and as such it sits near the top of the evidence hierarchy. A well performed, high quality experiment limits bias and controls for extraneous or confounding factors which may otherwise influence the results observed. This allows the conclusion to be made that indeed the intervention administered, in this case the drug to lower blood pressure, is in fact responsible for, or has 'caused', the recorded reduction in cardiovascular mortality in the patients compared to those individuals who were unfortunately randomly assigned to receive either a placebo or some other less effective treatment (but were otherwise treated identically within the same experiment). Synthesis of evidence informing the effectiveness of an intervention or therapy, particularly experimental evidence, is the focus of the fourth book in this series and will not be addressed in any specific detail in this volume beyond where necessary comparisons in study design between experimental and observational or epidemiological studies are made.

The Role of Evidence Informing Risk in Health care

Questions to inform practice are necessarily asked in health care that move beyond the effectiveness of interventions for prevention and/or treatment of a disease. If one is considering the etiology of an illness or health problem there are a number of reasons why an experiment, the gold standard for informing questions related to effectiveness, is unwarranted, impractical or in fact, unethical. To fully inform decision makers any synthesis of evidence related to the effectiveness of a treatment, or the beneficial outcomes due to the intervention being researched, to be balanced, should also inform as to the harms or risk of adverse outcomes due to the intervention administered or some exposure received. This evidence related to the adverse outcomes related to an intervention administered in health care practice can, and is to some extent, informed by RCTs. Beyond evidence related to the harm of an intervention or study, to establish or quantify the risk of a causative agent for a disease or health problem, observational studies are an essential source of evidence. The majority of studies published in clinical journals and accessible in medical databases, represent observational as opposed to experimental research (Egger et al., 2008).

The overarching term 'observational' studies, also called epidemiological studies, encompasses a broad range of study designs including cohort studies, case-control studies, cross

sectional studies, case series and case reports. These studies serve useful purposes and are characteristically the first scientific mark, beyond any anecdotal evidence, of a potential cause for a disease state. Investigators engaged in observational research may refer to existing data, in patient records for example, to confirm their scientific hunch or hypothesis. Conversely, researchers may opt to design a specific study to investigate a new idea and collect new data to add strength to any existing or new associations. Therefore, observational studies may serve a variety of purposes ranging from discovery of new associations to confirming existing associations or, potentially refuting previously reported associations between exposure and disease.

Irrespective of the intended aims of the study and the resulting conclusions, characteristically in observational studies, the investigator takes a less active role when compared to trials (Hoppe et al., 2009). Rather, these studies take advantage of the fact that people are exposed to any number of possibilities throughout their day-to-day life and are not limited to preventions and treatments that are commonly the focus of medical practice (Egger et al., 2008). Where a trial may refer to an 'intervention', synonymous with this concept in these studies is the term 'exposure'. Subjects being 'observed' by such research studies, are exposed either by their own choice, for example opting to indulge in a cigarette or a glass of wine habitually after dinner, or maybe 'exposed' inadvertently, such as living within the neighborhood of a wayward nuclear facility, or having received a particular vaccination to protect against influenza some 20 years or so previously when a child.

These observational, or epidemiological, population-based study designs are subject to biases and confounding infrequently encountered in RCTs and as such many critics believe they are of limited worth at informing health care or to enable any useful inference from the results collected (Black, 1996). To some degree, when informing evidence of effectiveness, this argument may carry some weight, however by the same token, when no trial evidence for the intervention in question exists, the best available evidence, in the form of observational studies, irrespective of inherent biases or confounding, should still be considered by clinicians in their decision making regarding both the treatment of an individual patient and the likely cause of the patient's condition, coupled with their own experience and expertise.

To inform the etiology of disease, a great deal of epidemiological research focuses on the establishment of associations between environmental risk factors, both modifiable and non modifiable, and the development of disease to establish a 'causal' link between the two (Blettner et al., 1999). This type of research is common and of great interest to both health care professionals and the lay public in general. Despite this interest, skepticism continually mounts as society is consistently inundated by the media reporting on the latest risks unearthed by researchers associated with the business of day-to-day life (Wenban, 2001). This research that is reported on is often observational as opposed to experimental, however this detail is rarely pointed out to the lay public – it is hard science delivered by experts – though preferably easily sensationalized! Results are reported with little regard for issues of quality or applicability of the research or any elaboration on the contradictory nature of the results with a media report on a similar exposure potentially only a few days previous.

Despite their potential to inform questions about etiology of disease, it is worth noting also that it is as a direct result of the lack of control of confounding factors and inherent biases that are unavoidable due to the nature of their study design, that there has similarly been protracted

debate regarding the use of observational studies to inform questions of effectiveness of therapies where the RCT is preferable (Black, 1996, Egger et al., 1998, Egger et al., 2008). Nutritionists for example will be well aware of the impracticalities of randomizing subjects to strict dietary regimes for prolonged periods. As such observational studies, in this case a prospective cohort study for example (see below), will be the most desirable option from which researchers and the media will invariably make claims as to the effectiveness of the dietary intervention and imply causal inference in the process.

When establishing the association or relationship between an exposure and health outcome, despite any debates, non-experimental research is essential. These studies are sometimes referred to also as correlational studies as they aim to summarize associations between variables, but are unable to make direct inferences about cause and effect as there are too many unknown factors that could potentially influence the data. Despite this admission, which must be acknowledged, studies addressing risk are conducted specifically to establish potential 'cause' of disease. Specific details of these unknown factors that preclude causal inference and how they relate to validity, bias and confounding will be addressed in a subsequent chapter of this volume. Here, where observational studies appear in the hierarchy of quantitative evidence will be discussed and details of unique features of their study designs explored.

The Quantitative evidence hierarchy and Types of Study design

In all considerations of quantitative evidence in the field of evidence based health care, a great deal of emphasis is placed on the 'type' of study design that has been employed by the research evidence and that the results in question are based upon. As mentioned, the topic of risk discussed in this volume rarely incorporates evidence from experimental studies (i.e. RCTs). However, when considering evidence based health care in general and the systematic review of any form of evidence, the full evidence hierarchy from systematic review through to cross sectional studies and case series, and the full range of variations in study design are worthy of some consideration here.

This 'ranking' of evidence based on study design is often referred to as the 'hierarchy' of evidence or an "evidence pyramid" (Figure 1). Construction and organization of this hierarchy is on the basis of the ability of the study design to minimize bias, and maximization of the internal validity possible within each type of study design. The observational study designs that inform questions of risk and etiology of disease are described in more detail below.

Systematic reviews

Systematic Reviews are found at the peak of the evidence pyramid. They take the form of an exhaustive search of the scientific literature, usually focusing on a specific aspect of a clinical topic, to identify all relevant and methodologically sound studies (Egger and Smith, 2008, Tricco et al., 2011). The studies are appraised and reviewed, and the results summarized. A systematic review of quantitative evidence may also contain a meta-analysis though this is not the defining feature of any systematic review (Loke et al., 2007, Tricco et al., 2011). In a meta-analysis, if a number of valid studies on a topic are sufficiently similar in one or more of the aspects of interest to the review question, the results of these studies can be combined and analyzed statistically as if they were from one large study (Egger and Smith, 2008).

Figure 1: The quantitative evidence hierarchy or pyramid
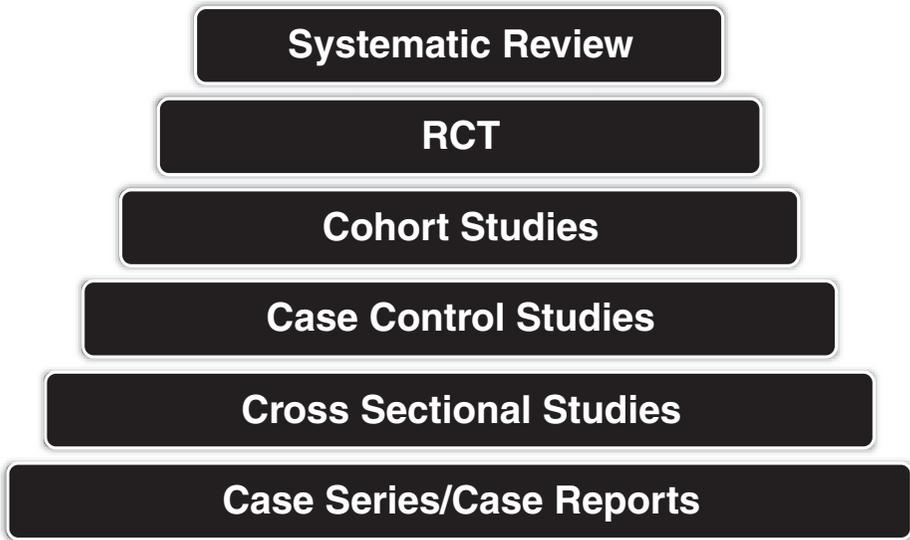
Randomised Controlled trial and Experimental study designs

In terms of primary research RCTs, well controlled experiments performed on subjects; represent the studies that carry the most weight when translating into evidence for practice. All aspects of RCTs are carefully planned to study the effect of a therapy (and 'only' the therapy or intervention of interest) on patients - this is done by comparing intervention and control groups. This study design necessarily includes methods that reduce the potential for bias (e.g. randomization and blinding where possible) (Roberts and Dicenso, 1999, Clancy, 2002). Though most experimental studies are directed towards informing the effectiveness of an intervention or therapy, they also, though not as frequently, have the capacity to provide valuable information related to harms or adverse effects from due to an intervention administered (Loke et al., 2007).

Cohort Study

Next in the hierarchy appears the cohort study, which along with the remaining study designs in the rest of the hierarchy of evidence (Figure 1), are classed as observational studies (Stephenson and Babiker, 2000, Hoppe et al., 2009). The cohort study is often referred to as the gold standard of observational epidemiological designs and is a form of longitudinal study that is commonly used to study exposure-disease associations. Cohort studies commonly investigate a large population of subjects over time that have an exposure, potentially a treatment, of interest and compare them with another group that has not been affected by, or treated with the exposure of interest (Stephenson and Babiker, 2000, Hoppe et al., 2009).

The term cohort, derived from the Latin military term referring to a unit of organization of soldiers within a Roman legion, is a group of people (subjects) who share a common characteristic, often within a defined period. Examples could be a group of people born in the same

year (birth cohort), or who may have been exposed to a drug or a vaccine. Here exposure may refer to a risk factor, prognostic factor or treatment with some intervention; surgical or pharmacological for example. As such these studies are useful to determine the incidence and natural history of a disorder or exposure. A cohort study may or may not involve a comparison group (Tay and Tinmouth, 2007, Hoppe et al., 2009). The comparison group may be the general population from which the cohort is drawn, or it may be another cohort of persons who have had little or no exposure to the substance under investigation, but otherwise similar. Groups like doctors, civil servants and surgery patients are often chosen as the source of the groups because they are relatively easy to define and/or monitor. Subgroups within a cohort may also be compared with each other. As individuals are not randomly assigned in cohort studies they are more prone to biases than RCTs (Tay and Tinmouth, 2007, Hoppe et al., 2009).

Cohort studies can come in two types, prospective and retrospective (Tay and Tinmouth, 2007). One of the defining features of a prospective cohort study is that the cohort is identified before the appearance of the disease under investigation and the outcomes of the group are then tracked forward and followed over time (Stephenson and Babiker, 2000, Tay and Tinmouth, 2007, Hoppe et al., 2009), sometimes for many years and across generations. A well known example of this is the Framingham heart study (Lloyd-Jones et al., 2002, Lloyd-Jones et al., 2004, Murabito et al., 1997). The cohort cannot therefore be defined as a group of people who already have the disease. The individual differences that exist in a population will therefore be represented in any sample of that population, for example amongst a sample, some people may smoke whilst others do not. The incidence rates for the disease under study are commonly determined in key subgroups. The study groups, so defined, are observed over a period of time to determine the frequency of new incidence of the studied disease among them (Stephenson and Babiker, 2000, Tay and Tinmouth, 2007, Hoppe et al., 2009). For example, in a sample that was free of lung cancer at the outset of the study, it may be anticipated that non-smokers will have the lowest incidence rate of the disease after 25-years, followed by moderate smokers, and that lung cancer will be most common in people classified as heavy smokers. The prospective study is important for research on the etiology of diseases and disorders in humans because, as mentioned, for ethical reasons people cannot be deliberately exposed to suspected risk factors in controlled experiments.

Prospective cohort studies investigating exposure and disease strongly aide in studying causal associations, though distinguishing true causality may well require corroboration from further experimental trials (Egger et al., 2008). The advantage of prospective cohort study data is the longitudinal observation of the individual through time, and the collection of data at regular intervals, so recall error is reduced (this is a recognized potential source for bias in retrospective studies) (Tay and Tinmouth, 2007). However, cohort studies are expensive to conduct, are sensitive to attrition and take a long follow-up time to generate useful data. Nevertheless, the results that are obtained from long-term cohort studies are of substantially superior quality to retrospective/cross-sectional studies, hence their place at the top of the hierarchy of observational study designs.

A retrospective cohort study is quite different from a prospective cohort study in the manner in which it is conducted. A retrospective cohort study, also called a historic cohort study, is a medical research study in which the medical records of groups of individuals who are alike

in many ways but differ by a certain characteristic (Tay and Tinmouth, 2007) - again, as an example, patients who smoke and those who do not smoke are compared for a particular outcome, such as the development of lung cancer. Effectively, in a retrospective cohort study, all the events - exposure, latent period, and subsequent development of disease have already occurred in the past (Stephenson and Babiker, 2000, Tay and Tinmouth, 2007). The data is simply collected now, and the risk of developing a disease established if subjects appear to have been exposed to a particular risk factor. There is no follow up of patients, as is the case with a prospective study. Clearly, a retrospective study has the benefits of being cheaper and less time consuming with resources mainly directed at data collection. Statistically, the two forms of the cohort study differ also. Whereas prospective cohorts should be summarized with the relative risk, retrospective cohorts, as with all retrospective study designs, should be summarized with the odds ratio.

A particular variation of a cohort study is a before and after study (or time series) (Tay and Tinmouth, 2007). In a before and after study, a measurement of interest is taken before and after a population sample is exposed to an intervention. Similarly, the nested case-control study can also be viewed as a subset of a cohort study. In a nested case control study, cases of a disease of interest which are identified within the bounds of a cohort study are focused on for the analysis and future study (Ernster, 1994). This design allows reductions in cost of data collection and analysis with only relatively minor statistical implications (Ernster, 1994).

Case-Control Study

Case-control studies represent studies in which patients who have an existing specific condition are compared with people who do not. These studies are also classed as 'observational' and therefore the subsequent correlational analysis usually presented from this type of research cannot conclude causality (Stephenson and Babiker, 2000, Hoppe et al., 2009). Case-control studies are often used as a rapid means of study of risk factors and therefore when considering evidence of risk of an outcome due to a prior exposure case-control studies are some of the most frequently occurring studies any systematic reviewer of the evidence will encounter.

Unlike the cohort study, case-control studies select patients who already have the disease or other condition of interest ('cases') and look back to see if there are characteristics of these patients that differ from those who don't have the disease ('controls') (Stephenson and Babiker, 2000, Hoppe et al., 2009). The main purpose of matching is to control for confounding (see below). Case-control studies are a relatively inexpensive to conduct and a frequently used type of epidemiological study that can be carried out by small teams or individual researchers in single facilities without the rigid structure of an experimental study (Hoppe et al., 2009). Despite the retrospective, non-randomized nature of a case-control study limiting the conclusions that can be drawn, this study design has claim to a number of important discoveries and advances in medical research, including the demonstration of the link between tobacco smoking and lung cancer (Doll, 2002).

It is worth noting also that the term retrospective study is sometimes used as another term for a case-control study (Hoppe et al., 2009). In the case-control study, the association is determined for each individual case-control pair, and then aggregated. This provides a more specific analysis of the possible associations, and potentially determines more accurately

which possible causes are directly related to the effect being studied, and which are merely related by a common cause. As with the cohort study above the retrospective design of the case control study should also provide a summary estimate of risk as an odds ratio.

Cross-sectional Studies

Cross-sectional studies can be thought of as providing a "snapshot" of the frequency and characteristics of a disease in a population at a particular point in time (Stephenson and Babiker, 2000). They are also referred to as prevalence studies and frequently employ survey research methods, which aim to find the same kind of relationships that might be shown over time in a cohort study, but at far less cost (Stephenson and Babiker, 2000, Kestenbaum, 2009). They are frequently used to assess the prevalence of acute or chronic conditions in a population. Cross-sectional research takes a 'slice' of its target group and bases its overall finding on the views or behaviors of the sample targeted, assuming them to be representative of the larger population. It is worth noting, since exposure and disease status are measured at the same point in time however, it may not always be possible to distinguish whether the exposure preceded or followed the disease (Kestenbaum, 2009, Stephenson and Babiker, 2000).

In a cross-sectional survey, a specific group is looked at to see if a substance or activity, say sedentary office work, is related to the health effect being investigated - for example, lower back pain. If a significantly greater number of office workers already have experienced low back pain than those workers who aren't sitting at a desk for most of the day, this would support the hypothesis that office work is correlated with low back pain. In epidemiology, cross-sectional studies often involve secondary analysis of data collected for another purpose. Data from these studies are useful in providing information about the health status and needs of a population (Stephenson and Babiker, 2000). Major sources of such data are often large institutions like a national government department responsible for collecting population information or census. Such studies can cover study groups as large as the entire population of the United States or groups from different countries around the world, but others are small and geographically limited.

Cross-sectional studies, which reveal clues to exposure/disease relationships, are often used as precursors to subsequent studies using more robust experimental design from higher up the evidence hierarchy to study a relationship, such as case-control, cohort studies or sometimes even RCTs.

Case series/Case report

Case series and case reports represent collections of reports on the treatment of individual patients or a report on a single patient respectively (Hoppe et al., 2009). It is difficult to portray valid statistics from such reports as they lack any controls and as such they are often used solely for description of conditions or observations subsequent to interventions or exposures.

A case series (also known as a clinical series) is a medical research study that tracks patients with a known exposure given similar treatment (prospective) or examines their medical records retrospectively for exposure and outcome, similar to a case-control study though generally with a smaller sample size (Hoppe et al., 2009). Case series may be confounded by selection bias, which limits statements on the causality of correlations observed.

In medicine, a case report is a detailed report of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient. Case reports may contain a demographic profile of the patient, but usually describe an unusual or novel occurrence (Hoppe et al., 2009, Vandenbroucke, 2001). A case report is a type of anecdotal evidence. Even more so than the other observational study designs described here, it is less scientifically rigorous than clinical data involving a larger sample size. Proponents argue that case reports have value within medical education and research in that they permit discovery of new diseases and unexpected effects (adverse or beneficial) as well as the study of mechanisms (Vandenbroucke, 2001). Case reports and case series have a high sensitivity for detecting novelty and therefore remain one of the cornerstones of medical progress; they provide many new ideas in medicine (Vandenbroucke, 2001).

At the base of the pyramid is 'bench' research in laboratories more often than not performed on animals. This is where most research into ideas, techniques, therapies and risk factors for disease starts.

Conclusion

Risk is a concept we are all familiar with. All of us, from all walks of life, are interested in risk and to varying extents weigh up our knowledge of risks in many of the decisions we make in our day to day lives, often unwittingly. It may be in simple decisions related to whether we decide to have an extra drink in the evening or to not use the seat belt in our motor vehicles whilst driving. We make such everyday decisions, frequently based on what we are informed the risks are related to certain behaviors or activities by the media, experts, policy makers and health care practitioners.

Questions related to risk in health care are commonly linked to attempting to uncover the cause, or etiology of disease. In an attempt to inform such questions researchers will rarely perform an experiment, despite the experimental study such as the RCT being the ideal study design to infer causation. This is because in most (but not all) cases pertaining to risk such an experiment would be unwarranted and unethical. More commonly, most research of this type comes from further down the evidence hierarchy and is observational, or epidemiological, in nature.

In observational studies, the researcher plays a much less active role than in an experimental study. Rather, they let nature take its course and quite literally 'observe' the effects. For this reason, where the term treatment or intervention is commonly used in trials, many observational studies will frequently refer to an 'exposure'. This is often to reflect the fact that sometimes subjects have no control over, or are even unaware of the exposure of interest. The gold standard of observational research is the prospective cohort study, though other study designs, particularly the case-control study also commonly inform such questions. Such studies are ideal to investigate rare diseases often with a long latency. The hierarchy itself is arranged based on the ability of the study design to limit risk of bias.

Evidence related to risk, derived from the range of observational and epidemiological study designs is important knowledge for health care professionals and policy makers. Just as with systematic review of RCTs, synthesis of such evidence better informs as to the relationships between exposure and outcomes and allows stronger conclusions and inferences to be made.

Chapter 2:

Validity, generalizability and quality in quantitative research

Scientific research seeks to establish relationships between two or more variables. All values reported in research reflect some 'measurement' of the variables under investigation. Statisticians will commonly use terms such as the *independent* variable, the variable that is manipulated or controlled by the researcher or often natural events in the case of observational research, and the *dependent* variable – the variable that is being measured. Clinicians will recognise the terms *intervention* or *exposure* and *outcome* respectively, as analogous descriptors. Two key characteristics of the relationships between variables investigated scientifically and consistently reported on are the strength or magnitude of the relationship, and the direction of the relationship.

When considering quantitative evidence describing relationships numerically, health care professionals will invariably encounter statistics and the mathematical analyses of probability that underlie them. Statistical analyses are necessary due to the uncertainty inherent in all of the measures we make in science - nothing in scientific research and the myriad of relationships it attempts to describe is certain. Statistics, as a science itself, is the almost universally accepted means by which we quantify, represent and describe the uncertainty, or lack thereof, in the relationships we investigate. Armed with this understanding however, it is important to note from the outset what we rather simply describe or define as 'quantitative research' does exactly as the name suggests – it generates numbers, which include the statistics – statistics, however, does not interpret the numbers for us! Interpretation of the numerical values generated is up to the researcher who conducted the study and those of us who have, for whatever reason, interest in the results and the interpretations made. The interpretation of the results of research, whilst seemingly simple if we are faced with, for example, a very large effect size and clear statistical significance, must also look beyond the numerical values to consideration of biases and the validity of the research under consideration and also, importantly, consideration of the clinical significance of the results presented irrespective of how pronounced the statistical description of the relationship may appear.

This section of this volume will focus on the first of these last two issues and discussion of biases and the importance of validity, in its various forms, in establishing what is good quality research and the basis for the inferences we make and our belief of them. There is marked variation, even within the realms of a single study design, for example considering just RCTs or even solely cross sectional studies, in the rigor with which studies are conducted. For the reviewer undertaking synthesis of existing, published research, this variation in the quality of studies included in the synthesis may impact on the reliability of the conclusions made from the process of synthesising the evidence.

Inherent to consideration of quality of primary studies is the concept of validity; in short we generally align establishing the validity of inferences about the relationships between variables with the truth or believability of these claims (Valentine, 2009). Put simply, validity concerns the

degree to which an account is accurate. Any 'threats to validity' – factors that might lead to incorrect inference – if not carefully considered, may in effect lead us to believe something that is not true. Important threats to validity in quantitative research include bias and confounding (Clancy, 2002, Rochon et al., 2005). The RCT experimental design exhibits most of its defining features, randomisation, allocation concealment and potential blinding for example, to limit the biases which may threaten validity and hence, when considering primary research designs, its pride of place at the top of the evidence pyramid (see Figure 1). Observational studies do not afford all, or sometimes any, of these features in many cases and as such, issues of bias and confounding are inherent to these designs (Clancy, 2002, Hoppe et al., 2009).

Validity and bias

An informative introduction to the concept of validity in scientific research is presented by Valentine (2009) who categorizes four important elements of study quality: internal validity, external validity, construct validity and statistical conclusion validity.

Internal Validity

In simple terms, internal validity refers to how 'good' a study is and how well it is conducted (Valentine, 2009). Practically, it refers to the extent that the difference in outcomes between comparison groups is due to the intervention or exposure, as opposed to anything else (Rochon et al., 2005). The better the study, the more convinced we are that the conclusions are warranted by the observations made. High internal validity is a hallmark of RCTs and experimental studies as results recorded in the ideal, controlled experimental environment allow reliable inference to be drawn regarding whether the intervention or exposure investigated caused the measured outcome (Valentine, 2009). When compared to experimental studies, observational studies, which, for example, are commonly used to establish risk of disease development following exposure to a risk factor, are viewed to lack internal validity. This is not to say that some of the techniques commonly associated with experimental designs cannot be implemented in observational designs. For example, in a prospective cohort study where subjects are being followed up in the future, the researchers who assess any outcome measures should, if possible, be blinded to exposure status to limit bias and increase validity (Hoppe et al., 2009).

Despite this, bias is an issue for many of these study types. For example, by definition a selection bias exists when subjects used in a study differ systematically from the population with the same condition that they are claimed to represent (Rochon et al., 2005). Taking large, random samples in some study designs including cross sectional studies alleviates selection bias. However, it is almost inherent to the case-control study design, which requires that both cases and controls are deliberately representative of the same population by the process of matching (Clancy, 2002). Another example of bias that threatens internal validity in retrospective designs such as the case control study is recall bias. Where these studies rely on taking a patient history in the present (as opposed to referring to relevant records if such exist), to establish the exposure variable which has resulted in the disease in participants, they are dependent on the patients or subjects recollections of events or exposures which occurred sometime in the past. Furthermore, such subjects are more than likely to have made their own associations as to what 'caused' their disease state and are more likely to, even

non-intentionally, have marginalized or omitted in their account, other potentially important exposures as a result (Clancy, 2002, Hoppe et al., 2009).

External Validity

External validity encapsulates similar meaning across the gamut of research designs and perspectives. Be it in RCTs, epidemiological designs and also qualitative research, external validity refers to the generalizability of the inferences made in one study, to other similar, yet not identical scenarios, to the one being considered (Valentine, 2009). An example of consideration of external validity may encompass the applicability of a similar exposure or intervention to a setting or population not directly studied by the research under consideration. How well do the results of this trial which established the effectiveness of this drug with subjects sampled from amongst those who present to an acute hospital apply to patients with a similar condition prescribed the same drug in the community setting? A well conducted comparable cohort study, distinguished primarily from an experimental study by the lack of randomisation in sampling from the population of interest and the level of control required in the experimental setting to maximise internal validity and the reliability of the cause – effect relationship initially hypothesised, potentially outscores the experimental design when considering external validity. Observation of the effect of the exposure or treatment in the real world setting as opposed to the contrived experimental environment allows inferences of etiology of disease and even effects and harms of treatment interventions to be more easily applied to other populations, settings or contexts than those directly studied.

Construct Validity

Construct validity most commonly refers to how well the study undertaken measured what it set out to measure and as such is generally focused on outcome measures (Valentine, 2009). For example, does a psychometric instrument purporting to measure quality of life encompass this construct adequately across the items incorporated in it? Any such instrument should, for example have items exposing validated measures of both physical and mental quality of life.

Statistical Conclusion Validity

Finally, is statistical conclusion validity which refers to the validity of statistical inferences inform-ing the strength of the relationship investigated (Valentine, 2009). The results of all statistical inference tests should be interpreted in consideration of the assumptions that underlie the selection and use of the statistical test ultimately employed in the study (Deeks et al., 2008). In general in scientific research, we draw conclusions based on statistical methods constrained by the assumptions of the statistical test we decide to use. In principle, what further distin-guishes observational studies from RCTs is the validity of some of these assumptions (Hoppe et al., 2009). Violation of these assumptions, for example using a paired analysis where data was not recorded from the same, nor matched, participants in the treatment or exposure comparison would impact on the validity of the inferences drawn as a result.

Confounding

Confounding is another important threat to the validity of studies. Confounding occurs when the apparent effect or process that is recorded or observed is, in reality, not the true effect,

but rather there are one or more other processes at play – potentially unknown processes - resulting from differences between comparison groups! (Clancy, 2002, Mamdani et al., 2005). This leaves any interpretation of the results potentially erroneous, with the risk established by a study between an exposure and disease state, or harmful outcome due to a treatment administered, potentially over- or under-estimating the size of the effect or even the direction of the relationship. In other words, subjects exposed to the factor under investigation may in fact differ in a number of aspects that are relevant to the risk of developing the disease in question (Clancy, 2002, Mamdani et al., 2005).

For example, a well conducted cohort study may show an increased risk of type 2 diabetes in subjects who are not physically active in their leisure time. The study may present statistically significant and precise results demonstrating the increase in type 2 diabetes and the researchers have drawn this plausible conclusion based on them. However, due to the absence of randomization of subjects to exposure groups, a luxury that is afforded an experimental design, and resultant susceptibility to selection bias, any number of other plausible explanations, which the study design cannot discount, could also account for the development of type 2 diabetes in some individuals and not others. An important confounder to consider in such a case would be the nature of the study subjects' occupations with respect to the physical activity they perform daily in the course of their work. As any conclusions about the outcome of interest due to the amount of leisure time activity in participants will clearly depend on their non-leisure time activity - this is an important confounder to consider.

Confounding appears in all types of observational studies and often there will be attempts by the researchers in the methods of the study reported to limit or account for confounding (Clancy, 2002). Sometimes this may occur in the design phase of the study where for example, a confounding factor may be 'standardized' in groups by deliberate selection of subjects based on levels of the confounder. This may occur in both cohort and case control designs. Most commonly however, confounding factors are addressed during data analysis with adjustments made for such 'confounding factors' in the estimation of the overall association (Blettner et al., 1999, Normand et al., 2005, Peters and Mengersen, 2008, Greenland, 1987). Statistical adjustment relies on measuring confounding factors and, in essence, correcting for them to remove or lessen their impact on the observed results. This is most commonly done by statistical modeling using regression techniques such as multiple linear or logistic regressions (Normand et al., 2005, Peters and Mengersen, 2008, Greenland, 1987). If a study has adjusted the analysis for a confounding factor it is important to be aware that if the factor either hasn't been, or cannot be measured with sufficient precision it will still remain a threat to the validity of the results (Egger et al., 1998). These same authors provide illuminating examples of studies where statistical associations between exposure and effect may appear causal, but are in fact, implausible (Egger et al., 1998).

Quality of Reporting

Whilst considering study quality or how 'good' a study is based on the techniques and methods used by the investigators involved, it also worth considering, and making the distinction between this discussion and any discussion about quality of reporting in primary studies (Valentine, 2009). Following a rigorous process of critical appraisal to establish risk of bias and the validity of the studies selected for inclusion in the review synthesis, attempts at

informing health care practice with the best available evidence uncovered by the stringent process of systematic review can often be stalled; not solely by the realization that many studies are of poor 'quality' and do not warrant inclusion in the final synthesis of the evidence on this basis, but also in many cases, that vital information to make informed judgment of study quality has been omitted! Sometimes this is by chance and sometimes by design, where methods employed haven't been described due to the word limit imposed by many scientific editorials for example. Almost all reviewers will encounter the scenario where they are at difficulty to make the decision between poor quality, poor reporting or in some instances, a combination of the two (Valentine, 2009).

In recognition of the reality and impact the poor reporting of research can have on an otherwise high quality study guidelines for reporting standards for a range of study designs have been developed by consensus. The Consolidated Standards of reporting of trials, or CONSORT statement (Schulz et al., 2010), offers those publishing experimental work a framework for providing the necessary information so consumers of the research evidence are in a position to make an informed decision regarding the validity of the study and the inferences made. Similarly, subsequent to the CONSORT statement and in response to the large amount of observational research available, and the questionable quality of reporting of much of this research, a group of methodologists, researchers and editors developed the *St*rengthening the *R*eporting of *OB*servational Studies in *E*pidemiology (STROBE) recommendations to improve the quality of reporting of observational studies and to facilitate the critical appraisal process necessary for any systematic review (Vandenbroucke et al., 2007, von Elm et al., 2007). The STROBE statement consists of a checklist of 22 items applicable to published observational studies (Vandenbroucke et al., 2007, von Elm et al., 2007). The STROBE statement, and the checklists of items that should be included in reports of observational studies in general, cohort, case-control and cross-sectional studies can be readily accessed at: http://www.strobe-statement.org/

It is important to note that the checklists developed are not tools to evaluate the quality of observational research, which as described above is underpinned by the differing facets of the validity of the study, but rather the reporting of observational research (Vandenbroucke et al., 2007). As clarity of reporting is a prerequisite to evaluation, application of the principles in the STROBE statement cannot help but facilitate evaluation of observational research for quality (Egger et al., 2007). In a recent systematic review a number of useful assessment tools available for assessing the quality of non-experimental research are identified (Sanderson et al., 2007).

Conclusion

To inform decisions in health care, the evidence upon which they are based must be valid, or in other words, believable. Unlike RCTs, observational studies cannot rely on random allocation of study participants to limit selection bias and confounding. Issues about the validity of observational research are what consistently throw into question any inferences of causality attempted by researchers and epidemiologists and similarly researchers undertaking systematic reviews of such research.

To interpret evidence of risk, indeed any quantitative research evidence, inference is drawn from the data measured and statistical results calculated. Interpretation of the values requires a

clear understanding of the inherent drawbacks and fortes of the research designs in question and how biases can impact on the validity of the results. Furthermore, interpretation and inference from results also relies on an understanding of the assumptions of the statistical tests employed to calculate the results. Although observational study designs may lack internal validity due to biases and confounding factors, due to the lack of an active role played by the researcher akin to any attempt to 'control' an experiment, these studies in many ways represent real world situations and as such reflect high external validity, beyond that normally associated with the restricted sampling required of RCTs.

On occasion a reviewer may find their quest for the best available evidence impeded not by the quality of the research study performed, but rather the quality of the reporting of the research. To reduce the instances of this occurring, standards for reporting of observational research have been published, similar to those available for trials.

Chapter 3:
The Synthesis of Evidence of Risk

When addressing a research question and analysis of existing evidence to inform the relationship between some exposure and likelihood of a particular health outcome, there are a number of accepted methods of evidence synthesis that apply to quantitative research and in particular observational studies. These methods include traditional narrative review, meta-analysis from published data, meta-analysis with individual data, and prospectively planned pooled analysis, and are all well described by Blettner et al., 1999. Of these available methods, the systematic review of published literature is the focus of this volume. The systematic review may include meta-analysis from published data which provides a new quantitative overall estimate of the effect measure of interest. Such informative quantification is possible only if the nature of the evidence located permits it. Issues and factors that will determine whether such a combination of individual studies is possible are the focus of this chapter. It may eventuate that, after careful and critical analysis of the evidence unearthed, that the nature of the data precludes any new quantitative analysis of the published results. In this case review authors may be constrained to presenting the synthesis solely as a narrative summary of the relevant published results.

Systematic review and Meta-analysis

In brief, the systematic review process, a formalised method of evidence synthesis, consists of the development of an *a priori* protocol and necessitates the formulation of a question, searching for the evidence to inform the question; selecting papers that meet the inclusion criteria; critically appraising papers to determine if, on the basis of methodological quality, they merit inclusion in the review; extracting data from the included studies; and "pooling" or synthesising data across the studies included (Figure 2) (Egger and Smith, 2008, Tricco et al., 2011). The specific details related to each of these 'steps' in the systematic review process will be addressed in turn in the remaining chapters of this volume in the context of both the protocol (Chapter 4) and the review itself (Section 2).

Secondary research that poses questions focused on etiology of disease, informed by the association between an exposure and health outcome, is increasingly being undertaken. This is apparent by the increasing amount of systematic reviews and meta-analyses of observational studies that now appear amongst the research literature (Stroup et al., 2000). Systematic reviews and meta-analyses of observational studies are as common as reviews of RCTs (Egger et al., 2008).

Thus far the preceding chapters of this book have introduced the defining features of these study designs and delved into issues of validity and quality of the primary research studies. In considering the synthesis of these types of studies, this chapter will introduce in more detail, both the potential and pitfalls unique to synthesis of observational study designs that are most commonly encountered to provide evidence of risk. Some comparison to synthesis of

Figure 2: The systematic review process

```
┌─────────────────────┐
│  Formulate PICO/PICo │
│       Question       │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│      Develop         │
│   Search Strategy    │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Searching for the   │
│      Evidence        │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│   Selecting Studies  │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Critically Appraising│
│       Studies        │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│   Data Extraction    │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│    Data Synthesis    │
└─────────────────────┘
```
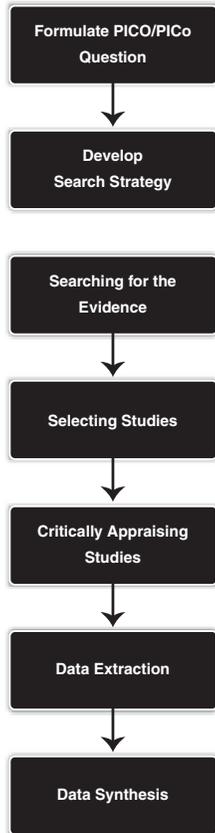
experimental studies, commonly used to inform health care practice, will be made throughout. The specific methods that are used will be addressed in a forthcoming section (Chapter 8) addressing briefly the meta-analytical methods used and encountered in the systematic review of such studies.

Meta-analysis is a quantitative method of combining the results of independent studies. This is achieved by pooling the results of various studies, in effect increasing the total sample size of the analysis, and improving the precision of the outcome estimate (Egger et al., 1997). Practitioners may be familiar with meta-analysis as it is commonly employed particularly where questions of the effectiveness of a therapy or intervention are principally being investigated. Such meta-analyses and the principles that statistical principles and assumptions have been well developed and utilized by the Cochrane Collaboration to inform the field of medical practice (Green et al., 2008), and to a lesser extent to date nursing and allied health, for a number of years.

In contrast to the use of meta-analysis in RCTs, the principle purpose of meta-analysis of observational studies is not to derive an overall estimate of effect, but rather to investigate the reasons for differences in risk estimates between studies and to discover patterns of risk among study results (Kheifets et al., 1997). This principle is important to appreciate and keep in mind from the outset when undertaking systematic review of evidence pertaining to risk, as it has important implications both for the type of analysis undertaken and its presentation (see Chapter 8). In other words, beyond calculation of a more precise overall measure of effect, the aims of such an undertaking include to assess whether an exposure is indeed a 'risk' factor for the health outcome or disease, to investigate differences or heterogeneity between individual studies, and also to investigate rare exposures and risks associated with rare outcomes or diseases (Blettner et al., 1999).

Limitations and obstacles in Meta-analyses of Observational studies

Inferences drawn from results of syntheses of observational research are subject to the same limitations and issues related to validity as do the primary research studies they draw together. Meta-analyses of observational studies and the overall risk estimate calculated may be affected by errors in measurement of the exposure variables, confounding and also biases, some of which were addressed earlier, that do not normally impact on a statistical combination of RCTs (Egger et al., 1998).

A further consideration or limitation that threatens the validity of meta-analyses of published papers, including those in epidemiology is that of publication bias (Blettner et al., 1999, Egger et al., 1998, Egger et al., 2008). As authors are more likely to publish significant results, particularly if unexpected, an overestimation of risk estimate is likely. Though publication bias has also been recognized as a limitation of systematic reviews of trials, concerted effort by the international research community has moved towards addressing this by establishing registries of all trials performed. No equivalent exists for epidemiological research.

Researchers undertaking a systematic review and potentially meta-analysis of evidence of risk which lends itself to observational epidemiological studies will rapidly appreciate difficulties that will arise that are unique to such an undertaking. These difficulties are due to differences in study design, differences in data collection methods and differences in definition of both exposure and confounder variables to name a few. Differences in the analysis and presentation of the data, no matter how similar the design of the studies, are particularly frustrating (Bekkering et al., 2008). Reviewers will often have to contend with instances where separate studies investigating the same association between exposure and disease or health outcome have provided different measures of risk too dissimilar to combine, or similarly, that adjust the analyses for different confounding factors (Bekkering et al., 2008, Peters and Mengersen, 2008).

These differences can have profound impact on the results of the research and the inferences drawn from them. Impact of differences in definition of exposure variables on a systematic review project can be highlighted with an example investigating the association between sedentary lifestyle and adverse health outcomes. At first glance the scope of such a review seems relatively straightforward. But what do we mean by sedentary? Conceptually, the term instantly imparts in most of us connotations of individuals who spend most of their

time sitting around and who don't move around much. The researcher who commits into such a review undertaking, armed with this clear understanding of 'sedentary' behavior will undoubtedly, probably around the stage of critical appraisal when study methods and results are teased apart in great detail, undergo an epiphany of sorts. So how is sedentary behavior measured? The sound conceptual understanding detailed above would lead the reviewer primed to encounter studies that have measured the time subjects spend sitting, be it during their leisure time, time spent at work, or both. Upon addressing the literature in the area it will rapidly dawn on the reviewer that sedentary behavior has two, rather distinct 'operational' definitions (Valentine, 2009), each reflected by how they are measured in practice (Owen et al., 2010, Pate et al., 2008). Yes, some research, as initially expected, does measure and report on the time subjects spent sitting per day or per week. Other research purportedly investigating sedentary behavior however, measures and reports on lack of physical activity – which doesn't necessarily imply subjects are sitting down – rather subjects may have been asked to simply report on the number of hours per week they spend doing low intensity tasks which may include sitting and watching TV or the like (Owen et al., 2010, Pate et al., 2008). Here the review designed to investigate the risk of exposure to a sedentary lifestyle has inadvertently discovered that this same concept has been approached and operationalized, with a different construct measured, in different ways by different studies (Valentine, 2009).

Furthermore, studies investigating the association between an exposure and disease, for example the associations between alcohol consumption and hepatic disorders, frequently look to strengthen the support for inference of a causal association between these variables by investigating a dose-response relationship between exposure and outcome (Bekkering et al., 2008, Blettner et al., 1999, Chêne and Thompson, 1996, Greenland and Longnecker, 1992). In such an example, reviewers would find that the primary research does not just report a comparison between a group of individuals who drink often, compared to those who drink little or no alcohol. Rather, these results would appear presented across stratified groups, or quantiles, based on the amount of alcohol consumption and the risk for each group (Bekkering et al., 2008, Chêne and Thompson, 1996, Greenland and Longnecker, 1992). One may observe a trend for the risk estimate to increase in individuals who fit categorical groups based on 1-2 glasses/week, 3-4 glasses/week or 5+ glasses per week. A similar study may present a similar trend in grams of alcohol consumed rather than by the glass. Across these same groups papers may report odds, risk or hazard ratios. Some, but not all, of these differences can be accounted for by a range of corrections, transformations and assumptions.

A recent study investigated how often these types of studies that report dose-response associations as a risk estimate per unit increase in exposure can be used in meta-analysis (Bekkering et al., 2008). The authors reported that only 61% of results from 71% of papers investigating the association between diet and prostate or bladder cancer were reported in a form usable in dose-response meta-analyses (Bekkering et al., 2008). Important missing information related to the levels of exposure was the main reason for omission. Many examples exist where extraction of data from published articles which present data in different, complex formats is prone to error. Many of these errors can be reduced or avoided completely in analyses where there is access to, or investigators make their primary data available (Blettner et al., 1999). The reviewer who is faced with the issue of missing important data for

meta-analysis may have to resort to contacting the authors of studies to fill in the gaps. Such missing details in reporting of results are an important, though sometimes unavoidable, threat to the validity of systematic reviews of this type of research (Bekkering et al., 2008).

Heterogeneity

Some of the differences encountered between studies, if large and diverse enough will preclude the prospects of any meta-analytical combination of individual study data entirely and the reviewer may find themselves limited solely to presenting a narrative synthesis of the evidence unearthed. Despite being able to statistically combine the results of the separate and diverse studies, differences between individual studies may leave the reviewer questioning whether they should combine such data (Egger et al., 1997, Egger et al., 2008). One of the underlying principles of meta-analysis is that studies should be 'homogeneous', or similar enough, to combine. Consideration of these issues raises the question of the appropriateness of pooling or combining the results of different studies together due to their differences, or the apparent heterogeneity between studies. Heterogeneity between studies is often more common and extreme in observational studies than clinical studies (Egger et al., 1997, Egger et al., 2008). There is a danger in combining heterogeneous studies together, particularly when they are from different populations or assess similar but inherently different exposures (Egger et al., 2008).

Heterogeneity pertinent to quantitative evidence synthesis can be classed into three main types: clinical, methodological and statistical (Deeks et al., 2008, Higgins et al., 2003). Clinical heterogeneity is due to identifiable 'clinical' differences between the included studies. This may include differences in participant characteristics such as age or sex, differences in exposures being investigated and their setting, duration or intensity, and differences in definitions of outcomes being measured for example (Deeks et al., 2008). Methodological heterogeneity is due to the differences between the methodologies of the studies conducted (Deeks et al., 2008). In meta-analyses of observational studies, it refers not only to the expansive differences in study design available to inform evidence of risk, but also differences between separate studies of the same design looking to be combined statistically. Such factors may include differences in study quality determined by the results of critical appraisal for example, or differences in the confounding factors measured and adjusted for. Statistical heterogeneity is an important consideration in practice confirming or questioning the decision to undertake the statistical combination of the data (Deeks et al., 2008). Dependent on the analysis executed there are a variety of ways to explore and examine statistical heterogeneity ranging from visual inspection of the results to other statistical procedures and operations to quantify, mathematically, the presence of heterogeneity (Higgins et al., 2003).

The MOOSE Statement

Similar to the aims of the STROBE statement (von Elm et al., 2007) in relation to the reporting of observational studies, in response to the substantial increase in published meta-analyses of observational research studies the *M*eta-analysis Of Observational Studies in Epidemiology (MOOSE) group developed a checklist of items for reporting of these meta-analyses based on a systematic review of the literature (Stroup et al., 2000). The proposed checklist contains

specifications for reporting of meta-analyses of observational studies in epidemiology intended to improve the usefulness of meta-analyses for all interested parties.

Conclusion

The systematic review of observational studies to inform questions of risk is becoming a more frequent practice amongst researchers. Systematic review of observational research is undertaken, as with reviews of effectiveness, to provide a more precise overall summary estimate of effect and more convincing evidence of associations between exposure and health outcome.

Systematic review of observational studies is fraught with many of the obstacles to inference that plague the primary research they look to synthesize. Selection bias and other confounding factors arising due to lack of randomization and experimental control may lead to observable heterogeneity in meta-analyses. Heterogeneity may arise due to methodological, clinical or statistical reasons and if present, may preclude statistical combination of data and leave the review author simply presenting a narrative synthesis of the available evidence related to the question of interest. The diverse range of formats in which outcome data may be presented, particularly as a result of differences in measurement of the exposure variable, can also be an impediment to meta-analysis of data derived from observational studies to inform questions of risk.

Irrespective of these obstacles, when addressing questions assessing the impact of weak or common exposures with potentially enormous implications for public health, such as the impact of passive smoking, the increased use of mobile phones or Wi-fi technologies or addition of additives to the food we purchase at the local supermarket, any systematic review may play an important part to lend weight to, or debunk, claims of causal links to disease and harm.

Chapter 4:
Developing a protocol

An *a priori* systematic review protocol is important because it pre-defines the objectives and methods of the systematic review. Furthermore, by stipulating the systematic approach to the conduct and report of the review to come it ensures transparency of process, which in turn allows the reader of the final review to see how the findings and any recommendations were arrived at. The protocol details the criteria the reviewers will use to include and exclude studies, to identify what data is important and how it will be extracted and synthesized.

A protocol provides the plan or proposal for the systematic review and as such is important in maintaining rigor (Tricco et al., 2011). Despite the best efforts of the review team, instances may arise during the course of the review project where the review deviates from what was initially described in the protocol. In these instances, any deviations between the protocol and systematic review report should be discussed in the systematic review report with clear indication and explanation for the reader of the reasoning behind them.

Review title

The title of the protocol should be as descriptive as is reasonable and reflect the type systematic review to be undertaken. Although a range of mnemonics have been described for different types of review (and research) questions, if, for example the review aims to examine risk this should, as much as possible, be stated clearly in the title of the document. If specific exposures and/or patient outcomes are to be examined these should also be included in the title. *For example:*

   "Smoking and risk or chronic obstructive pulmonary disease: a systematic review".

This example provides potential readers of the review with a clear indication of the exposure, and the outcome of interest, as well as the fact that it is a systematic review. Where possible the setting and target population should also be stated. A clear title is important for indexing and to assist peer reviewers as well as end users to identify the scope and relevance of the review.

Question development

Developing a good review question is the critical first step in undertaking a sound systematic review. The review question structures and guides key components of the entire review project: objectives, inclusion/exclusion criteria, the search strategy, the relevant data to extract and if defined specifically enough, may even lessen the likelihood of heterogeneity in any meta-analysis or statistical combination of the results of the individual studies performed. The PICO mnemonic, Population, Intervention, Comparator, Outcome, is a useful tool in breaking down and thinking through the components of a review idea or topic area (Table 1) (Stone, 2002).

Table 1. PICO Model of question development

P	Patient, Population, or defining feature (e.g. problem)	How would I describe a group of patients similar to mine?
I	Intervention, or Exposure	Which main intervention or exposure am I considering?
C	Comparison to Intervention (if appropriate)	What is the main alternative to compare with the intervention?
O	Outcome you would like to measure or achieve	What can I hope to accomplish, measure, or affect?
S	Type of study you want to find	What would be the best study design/methodology?

Clinicians well versed in practices and competencies which are the hallmark of an evidence based practitioner will recognize the mnemonic as an important first step in framing a question in an 'answerable format' to facilitate effective searching for evidence to inform decision making in practice (Dawes et al., 2005).

The PICO model is pragmatically suited to questions of therapy or diagnosis. Questions related to harm also suit the model well, as evidence to inform harm or iatrogenic disease can also be gleaned from trial data, similar to beneficial outcomes also described in terms of risk (Chou and Helfand, 2005). Review questions addressing etiology or investigating if a correlational association exists between variables may not fit the PICO model of question development so well. This does not preclude its use however, as in many cases this is simply due to the lack of a predefined comparator as may often be the case when considering questions of comparative effectiveness. More often than not the comparator is implicit - a lack of exposure to the variable or risk factor of interest and as such is rarely defined. Similarly, the single outcome of interest is also predefined in many cases, being the disease state of interest. Most consideration is frequently directed towards defining both the population and particularly the exposure variable of interest.

Other mnemonics which further expand the detail stipulated in the PICO are also expounded in the literature informing clinicians and researchers in question development. One such mnemonic is PICOS which also includes specific detail of the study design that is appropriate to be addressed for the review question being asked (Tricco et al., 2011). A more recent mnemonic, PICOTT, includes the 'type' of question being asked (T) as well as the 'type' of study design (T) (Schardt et al., 2007). Questions related to the effectiveness of a therapy or intervention often limit themselves solely to experimental studies as these, at the top of the evidence hierarchy, are the best available evidence to establish a causal relationship between intervention and outcome. A review of effectiveness will often not consider any other type of study design beyond experimental and preferably RCTs at that. Whilst establishing the best available evidence related to the risk attributable to an exposure, many reviews may limit themselves to prospective study designs only (Hemingway and Marmot, 1999, Renehan et al., 2008), as if well conducted they have less susceptibility to biases, whilst others may include information extracted from the full range of observational designs.

Many reviewers will attest to the fact that extra time spent initially conceiving and considering a well formulated, clear and explicit question has ultimately saved time and lessened subsequent reviewer consternation over the course of the review project. Difficulties invariably arise as a result of unclear or ambiguous direction arising when proper attention to components of the review question, which is ultimately unavoidable during study selection and appraisal, unearths apparent inconsistencies.

In addition to the PICO elements of the review question it is important that the review team be clear on the type of question being asked. Simply appreciating that the question lends itself to numerical data is rarely enough insight. For example is there clear distinction that is aiming to inform a question related to the harm of a surgical intervention or adverse drug event or other iatrogenic cause as opposed to one aiming to establish the relationship between a modifiable risk factor and disease prevalence in the community? Both of these questions, despite both looking to establish risk due to an exposure, will deviate in the appropriateness of populations included, the subtleties defining the exposures, whether a comparator is appropriate or indeed necessary, the outcomes determined to be the most relevant to answer the question, and also the predominant study design that represents the best available evidence to inform the strength of the relationship between variables and any inferences of causality.

Clear understanding of the question from the outset will also allow more informed selection of the instruments appropriate for critical appraisal of the included studies and to preempt the outcome measure used to describe the overall effect measure calculated in meta-analysis or if the analysis will likely necessitate multiple subgroups. All of these details, as much as is practicable, should be stipulated in the *a priori* protocol of the review. In the absence of clear understanding of the question being asked and the approach to the synthesis of the evidence the question demands, important considerations which could, and inevitably should, appear in the protocol will be omitted and leave the review team ill informed, and also at greater risk of not completing the review undertaking.

Background

As the name of the section implies, the background should describe important aspects of the issue under review, including the nature of the exposure or intervention, the target population it applies to, and the outcomes selected to be reported on. The background should provide sufficient detail to create a clear picture for the reader on each of these elements to justify the conduct of the review and the decisions about the choice of the various elements that the review intends to focus on.

To aid in creating this 'clear picture' for readers and users of the review, and also for the reviewers themselves throughout the project, the background section of the protocol provides the opportunity to explore and establish the important conceptual definitions which underpin the elements of the review question, and importantly, how, for the purposes of the review in question these concepts will be operationalized, or in more pragmatic terms, how they will be measured (Valentine, 2009). Despite similar concepts being researched and reviewed by different researchers globally, there may be marked differences in how researchers, undertaking both primary and secondary research, may define and measure these concepts. In the example described in the preceding chapter of the impact differences in the definition of

sedentary behavior used by different studies, some of this could be overcome by using specific objective criteria to aid in defining important concepts. In this case, measure of metabolic energy expenditure (MET) values of <1.5 would be indicative of sedentary behavior defined by a standardized level of low energy expenditure associated with various activities (Owen et al., 2010, Pate et al., 2008).

Dependent on the complexity of the issues involved and the current understanding in the field of inquiry, some aspects of the PICO elements may warrant greater focus in the background of the protocol and review such as the exposure or risk factor – such as in the example of sedentary behavior provided above. As another example, despite our universal conceptual understanding of pain, some extra attention may be afforded in the background by the review authors to support the use of only a selection of psychometric instruments, based on the established validity of their findings rather than including studies that measured pain irrespective of the instrument used. In this regard it is also often valuable to justify why elements are not to be included.

Systematic reviewers place significant emphasis on a comprehensive, clear and meaningful background section to every systematic review, particularly given the international circulation of systematic reviews, variation in local understandings of clinical practice, health service management and client /patient preferences and experiences. The background of the review protocol will also commonly conclude with some indication of other similar reviews (but not identical!) that have been published already.

Review Objective(s)

The objective(s) of the review should provide further qualifying statements of the presented title and the question(s) being addressed. Clarity of the objective(s) is achieved by specific reference to the identifiable elements of the PICO - population, intervention, comparator and outcomes. Beyond the advantages of developing a well formulated review question for guiding the entire review project visited briefly above, clearly stated objectives highlighting the PICO elements allow the protocol to be more effectively indexed, aid in developing the search strategy, and assist in focusing the protocol; any reader will have an instant appreciation of the scope and focus of the review even before addressing the detailed inclusion and exclusion criteria. Conventionally, a statement of the overall objective is made and elements of the review are then listed as review questions.

Review objective

To systematically review risk associated with second-hand cigarette smoke exposure on the exacerbation of dyspnea and physiologic deterioration of chronic obstructive pulmonary disease.

Review question

What is the effect of sustained, indoor second-hand cigarette smoke exposure versus no second-hand cigarette smoke on the exacerbation of dyspnea and pulmonary function in middle-aged to older adults with chronic obstructive pulmonary disease (COPD)?

Criteria for inclusion/exclusion

Once the researcher formulates the review question, the next step is to further specify each element of the question by establishing the inclusion and exclusion criteria. These criteria further define the scope of the review by clarifying the characteristics of studies the reviewer will consider for inclusion in the review.

Population

The reviewer specifies the characteristics of the population in the protocol and the review report to establish the breadth and applicability of the review findings. Typically, the reviewer needs to consider age, ethnicity, gender, presence and degree of existing conditions, disease trajectory and environment or setting. The reviewer must strike a balance between determining the attributes of the population that provide a clear focus for the review while maintaining a significant scope.

The inclusion and exclusion criteria need to reflect sound clinical and scientific reasoning and the need for an adequate degree of homogeneity amongst the samples in the candidate studies. For example, a very broad population may be all adults with obstructive lung disease that includes all specific conditions (emphysema, chronic bronchitis, bronchiectasis, asthma, etc), ages, and severity of illness. Alternatively, the question scope could be more restrictive if the reviewer specifies moderate-to-severe chronic bronchitis and emphysema patients older than 55 years in a stable phase of disease.

The reviewer makes the decisions about the inclusion criteria much like a primary researcher makes decisions about the population of interest while building a hypothesis or specific aims. Unlike the primary research context, the systematic reviewer needs to consider how other investigators typically examine the population of interest while selecting a population that is scientifically appropriate for the interventions and outcomes.

If an international standard exists for the diagnosis of a clinical problem, the reviewer should consider specifying the standard and its parameters. Take caution that newly developed diagnostic criteria may unnecessarily limit the eligibility of important studies (O'Connor et al., 2008). However, if a standard is long standing or easily interpreted in candidate trials, its use will help clarify the population to whom the reviewer expects to apply the findings. For example, in the case of patients with chronic obstructive pulmonary disease, the diagnostic inclusion criteria for the population might read:

> *Candidate studies will be considered if the sample represents middle-aged to older adult patients (greater than 45 years) with stable moderate-to-severe chronic obstructive pulmonary disease. The review will use the Global Initiative for Chronic Obstructive Lung Disease (GOLD) and the American Thoracic/European Respiratory Society Guidelines (ATS/ERS) diagnostic categories. The description of the severity of disease is as follows: stage II or moderate disease is a forced expiratory volume in one second (FEV_1) of 50-80% predicted; stage III or severe is an FEV_1 of 30-50% predicted and stage IV or very severe is an FEV_1 <30% predicted. COPD includes patients with chronic bronchitis and emphysema but not asthma so there should be no more than 15% improvement in FEV_1 after short-acting bronchodilator.*

Intervention/Exposure

When assessing risk, the review question will frequently refer to the exposure rather than an intervention. Like the parameters that define population, the inclusion criteria related to the intervention or exposure will determine scope and some degree of homogeneity in candidate studies. The intervention or exposure should be defined in a way that is specific enough to be logical and allows sound comparisons. For example, a review about the effect on nurses of interruptions and distractions during medication administration on error rates, could conceivably include a fairly wide range of exposure categories from environmental sources (noises, visual distractions), people (phone calls, conversation, pages, text messages, patient calls), care (urgent needs), and gaps (missing medications, misplaced medications). While a missing medication is quite different than a phone call, both cause interruption in the process of medication administration. The reviewer could effectively argue that it is legitimate to pool the effects of these various sources because they all result in an interruption.

Consider some of these issues while specifying the intervention or exposure:

- How does the exposure vary naturally or how does the intervention vary in its administration?
- Is there a need to specify the "density" of the exposure/intervention meaning the frequency, magnitude or dose, duration, timing, or source/person?
- Might the exposure or intervention be combined with other variables? If so, how will you sort the effect of the variable of interest from other influences?
- Are intervention/exposure subgroups distinguishable and relevant?

Like primary research, plan the analysis during the protocol to avoid statistical "fishing". So if the intervention/exposure lends itself to subgroup analysis (e.g., mild vs heavy exposure), specify them in the protocol.

In the example of the exposure to second-hand smoke and dyspnea and pulmonary function effects, the following exposure inclusion criteria may be reasonable:

The review will consider studies that examined the risk posed by direct, daily exposure to second-hand smoke in an indoor space lasting for several hours a day for at least a week or longer. Studies will be excluded if the participants were also current smokers. If the studies categorized heavy exposure from light exposure, subgroup analysis will be conducted.

Comparator

If the review question specifies a comparison exposure for head-to-head comparison, the scope of the review will be narrower and point to particular research designs that include a second group. In reviews of the effectiveness of an intervention, the comparator may be a placebo or sham treatment or a contrasting treatment. In risk assessment the comparator may be no exposure at all, a different degree of the same exposure, some other parallel exposure or it may not be defined. In the example of interruptions and distractions, a comparator might be a protected quiet environment or the comparison may not be specified. If a parallel exposure is planned, similar issues should be considered as those listed under exposure.

In the second-hand smoking example, the comparator might be specified as:

The review will consider studies where the adult patient with COPD had no indoor exposure to second-hand smoke.

Outcome

The review protocol must specify the important outcomes of interest relevant to the health issue and important to key stakeholders like the knowledge users, consumers, policy makers, payers and the like. Exposure to risk implies that the outcomes of interest will be adverse but it is worthwhile to consider both sides of the issue (benefit and harm) if possible, when designing the protocol. In addition, specify the outcomes to exclude.

Generally, reviewers should avoid surrogate outcomes as they may mislead. On the other hand, sometimes direct measurements may not be possible but bring additional caution to the interpretation. Consider when and how the outcome may be measured. In addition, determine if the review should examine secondary or mediating outcomes. For example, in the case of exposure to interruptions and distractions, the important outcome is any error in medication administration. This includes wrong dose, route, medication, person, timing, etc. Think about the outcomes as they might form a table of results.

In the worked example of the exposure to second-hand smoke, where the primary outcomes of interest are dyspnea and pulmonary function, the inclusion criteria might be:

- Any self-reported measure of dyspnea, breathlessness or breathing effort measured during or after exposure to second-hand smoke
- Pulmonary function tests that include FEV_1, FEV_1/FVC, total lung capacity (TLC), residual volume (RV)/TLC or other lung volumes relevant to exacerbations of COPD

Types of studies

The protocol should specify the types of study designs that align with the review questions. Questions relevant to exposure and risk outcomes will most frequently be answered with cohort (prospective and/or retrospective), cohort with comparisons, case-control studies, case series or cross-sectional studies. The protocol should specify the preferred design that is least vulnerable to significant sources of bias but recognize that in the absence of stronger designs, other study types will be considered. For example, if the exposure variable cannot be ethically manipulated (like second hand smoke), the strongest design to determine the effects of the exposure would be a prospective comparative longitudinal design. However, because these designs are fraught with difficulty, the reviewer would transparently identify other, albeit less robust, feasible designs worth considering for the review process.

In the example of the second-hand smoke and exacerbations of dyspnea and pulmonary function effects in patients with COPD, the following designs would be consistent with the question:

- *Prospective cohort-control*
- *Case-control studies*

In the absence of cohort-control or case-control studies, other designs will be considered including single group cohort designs, retrospective cohort studies, cross-sectional descriptive studies and case series will be considered for a narrative synthesis.

Search strategy

Systematic reviews are international sources of evidence; particular nuances of local context should be informed by and balanced against the best available international evidence. The protocol should provide a detailed strategy that will be used to identify all relevant international research within an agreed time frame. This should include databases that will be searched, and the search terms that will be used. In addition to this, it should also specify what research methods/methodologies will be considered for inclusion in the review (e.g. randomized controlled trials, cohort studies). Quantitative systematic reviews will often include a hierarchy of studies that will be considered, whereby the reviewer specifically states that they will be searching for randomised controlled trials due to their methodological rigor; however, in their absence or if there is only a small amount of low quality trials, other quantitative study designs, such as cohort or case controlled studies, will also be searched for and included. Comparatively, the reviewer may decide to be stricter in their search strategy, and state that they will only prospective cohort studies for example, and not include studies of lower quality. The risk here is that no studies of a pre-specified type will be identified, particularly for questions relating to evidence of risk, leading to a systematic review with no included studies and therefore unable to produce implications for practice. For some questions, there may be no research of any study design conducted on the topic, and text or expert opinion may thus be considered for inclusion in the review, as it constitutes the 'best available' evidence.

Within reviews the search strategy is described as a three-phase process, beginning with the identification of initial key words followed by analysis of the text words contained in the title and abstract, and of the index terms used in a bibliographic database to describe relevant articles. The second phase is to construct database-specific searches for each database included in the protocol, and the third phase is to review the reference lists of all studies that are retrieved for appraisal to search for additional studies. The search strategy may also include hand searching, which is where key journals in the field may be browsed to identify any relevant articles. Another strategy, particularly where literature is sparse, is to contact experts in the field who may be able to provide relevant study details or additional keywords.

The search strategy should also describe any limitations to the scope of searching in terms of dates, resources to be accessed or languages. Each of these may vary depending on the nature of the topic being reviewed, or the resources available to those conducting the review. Limiting the search by date may be used where the focus of the review is on a more recent intervention or innovation, however, potentially relevant studies as well as seminal, early studies in the field may be excluded and should thus be used with caution, the decision preferably to be endorsed by topic experts and justified in the protocol. The comprehensiveness of searching and the documentation of the databases searched is a core component of the systematic review's credibility. In addition to databases of commercially published research, there are several online sources of grey or unpublished literature that should be considered.

Assessment criteria

There are a variety of checklists and tools available to assess studies. Most checklists use a series of criteria that can be scored as being met, not met or unclear or not applicable. The decision as to whether or not to include a study can be made based on meeting a pre-determined proportion of all criteria, or on certain criteria being met. It is also possible to

weight the different criteria differently. Often differences in these choices made by the reviewer will depend on whether a checklist or a scale has been selected as the instrument of choice in quality appraisal (Shamliyan et al., 2010). These issues with an example of a checklist that may be used to aid the appraisal of cohort and case-control studies are discussed in more detail in Chapter 6.

In practice, as a minimum, to limit the impact of bias in a systematic review appraisal of each study should be done in duplicate without conferral between reviewers until both appraisals have been completed. It will often require discussion or sometimes the introduction of a third opinion to arrive at consensus regarding items in appraisal particularly if the result determines whether the paper will be included or excluded on the basis of quality. Provision for these processes should be detailed in the protocol.

Data extraction

Data extraction refers to the process of sourcing and recording relevant results from the original (or primary) research studies that have met the inclusion and appraisal criteria and are included in the systematic review. It is important that both reviewers use a standard extraction tool that has been piloted and is applied consistently. The protocol must therefore describe how data will be extracted and include an appropriate data extraction instrument in appendices to the protocol.

The process of data extraction should glean all of the relevant information from the included studies; this includes important details related to the sample demographics, diagnostic criteria for presence of a disease for example, or important details related to the exposure of interest such as the duration or level of exposure. Of primary importance during data extraction however is the recording of the appropriate outcome data. Reviewers must be clear regarding the outcome data they are looking to extract. This should be guided by the review question, not solely the data reported by the research paper. Unlike a reviewer addressing trial data, where for the majority of studies only a small proportion of the outcomes reported will eventually be extracted, most of the data presented by a study investigating etiology of disease will be relevant to the review question. As an example, in terms of outcome data a review investigating the risk of a brain tumor developing due to mobile phone use will present incidence of cancer as its primary outcome – what may differ and will have to be taken into consideration will be the frequency of exposure to individuals.

To effectively extract the appropriate data reviewers must also be intimate with the form the values they are searching for will take. Attempting to inform the risk or probability of a disease developing due to an exposure requires reviewers to understand, often dependent on the study design that is being addressed, the diverse ways in which association between variables can be described. Such data related to etiology of disease or health outcomes is generally dichotomous in nature, though may also be reported across ordinal scales aligned to different levels of the exposure variable. Risk ratios, relative risks, odds ratios, hazard ratios and the like, all represent measures of risk which may be encountered. Appreciation of the techniques that will be used in any subsequent synthesis or meta-analysis of the data is also recommended to ensure reviewers extract all of the necessary data. In many cases if all that can be extracted is an estimate such as a relative risk rather than the incidence data used to establish it (due to this not being reported) often error terms such as the standard error

or confidence limits will also be necessary to provide the complete set of data for synthesis (Greenland, 1987).

To address some of these issues of note in the process of extraction, there are recommended strategies to minimise the risk of error or omission when extracting data from studies. These include utilising a standardised data extraction form for the review project. Such a form can be tailored to prompt and direct reviewers to extract the appropriate and complete data with careful consideration in their development. Any instrument or form used for extraction should be tested in a pilot extraction prior to commencement of review to establish if users find them intuitive and they do lead to extraction of appropriate information and data. Training and assessment of data extractors can also increase the efficiency of the exercise. Importantly, as with the assessment of study quality that precedes any extraction of data from studies, it is preferable to have two people extract data from each study to limit error. Where important data is missing form studies provision is often made in the methods to attempt to contact the original authors of studies to request data.

Data synthesis

Just as the primary researcher describes data analysis in a research proposal, the systematic reviewer forecasts the plan for data synthesis in the protocol. Typically, reviewers specify the methods for both a statistical approach to synthesis via meta-analysis and a narrative summary should a meta-analysis be inappropriate. If a meta-analysis is anticipated, the reviewer would delineate the following in the protocol:

- A method to determine statistical heterogeneity between or amongst the observational studies
- The specific methods of pooling data across studies to determine risk or a method to compare risk between or amongst exposures
- Specific statistical methods to test for significance, the direction and size of the effect of exposure appropriate for the level of measurement and number of groups.
- Plans to conduct subgroup and or sensitivity analysis
- Confidence interval probability (eg, 95% confidence interval)
- Methods to insure reliability of data input (dual entry)

The second-hand smoking in COPD review might specify the data analysis like this:

Studies will, where possible, be pooled in statistical meta-analysis. All results will be subject to double data entry. Odds ratio (for categorical data) and weighted mean differences (for continuous data) and their 95% confidence intervals will be calculated for analysis using a random effects model. Heterogeneity will be assessed using the standard Chi-square test. Planned subgroup analysis will include mild and heavy exposure to second hand smoke. Where statistical pooling is not possible the findings will be presented in narrative form.

Narrative Summary

The protocol should also describe a process for developing a narrative summary to anticipate the possibility that meta-analysis may not be possible. In addition, all systematic reviews,

regardless of the ability to use summary statistics, require a summary of the included research. The narrative summary should draw upon the data extraction, with an emphasis on the textual summation of study characteristics and patterns as well as data relevant to the exposure and outcomes of interest. All systematic reviews of quantitative evidence should include a data table illustrating the key aspects (exposure, study design, sample design and size, results, etc.) of each included study. In addition, the reviewer must use appropriate judgment to determine how much any particular study and the pooled evidence contribute to synthetic, bottom-line recommendations.

Conflict of Interest

A statement should be included in every review protocol that either declares the absence of any conflict of interest, or describes a specified or potential conflict of interest. Conflict of interest statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje. org/ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) have extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Acknowledgements

The source of financial grants and other funding must be acknowledged, including the reviewers' commercial links and affiliations. The contribution of colleagues or Institutions should also be acknowledged.

References

Protocols most often use Vancouver style referencing. References should be numbered in the order in which they appear with superscript Arabic numerals in the order in which they appear in text. Full reference details should be listed in numerical order in the reference section.

More information about the Vancouver style is detailed in the International Committee of Medical Journal Editors' revised 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication', and can be found at http://www.ICMJE.org/.

Appendices

Appendices should be placed at the end of the protocol and be numbered with Roman numerals in the order in which they appear in text. At a minimum this will include critical appraisal and data extraction tools.

2

Conducting a Systematic Review of Evidence of Risk

Chapter 5:
Searching for Evidence

The comprehensiveness of searching and the documentation of the databases searched is a core component of the credibility of any systematic review. In addition to databases of published research, there are several online sources of grey or unpublished literature that should be considered. Grey literature is a term that refers to papers, reports, technical notes and other documents produced and published by government agencies, academic institutions and other groups that are not distributed or indexed by commercial publishers. Many of these documents are difficult to locate and obtain. Rather than compete with the published literature, grey literature has the potential to compliment and communicate findings to a wider audience.

As systematic reviews can be viewed as a research study of themselves, the methods used need to be transparent to the reader to the extent that the review can be replicated. Essentially, the searching phase of a systematic review can be seen as corresponding to the data collection stage of a research study, and therefore it is imperative that the methods used to search and retrieve studies is recorded and documented in detail. Reference management software, such as Endnote and RefWorks, can be useful tools to utilize during the searching process to track and record search results. However, other important information, such as date of searching, results retrieved from each database, specific strategies used for each database, should be recorded. Although common key terms will be utilized across databases, often the exact search strategy will differ, due to the different controlled vocabulary thesaurus or indexing systems the databases use; for example, Medline utilizes MeSH (Medical Subject Headings) as its thesaurus, whereas EMBASE utilizes unique EMTREE subject headings. These differences will lead to variations in the exact search strategy used for each database, and should be recorded. Most databases currently have systems where searches can be stored within the database itself, and rerun when needed. This facility may be useful to use when considering an update of a previously conducted systematic review. Furthermore, when developing your search strategies, you may wish to consult a librarian or health care informatics specialist if available to you.

Once a focused research question has been determined, relevant databases should be searched. Occasionally, you may wish to search for literature that specifically employs

observational research methods as opposed to all forms of published research applicable to your question of interest.

The methods used for searching need to be clearly recorded, so that they can be recorded and reported in line with the MOOSE statement (Stroup et al., 2000). The MOOSE statement recommends that when reporting the methods for searching used, the following needs to be included:

- databases named,
- the qualifications of those conducting the search stated,
- index terms used and search strategy
- Search software used and features of the software (such as explosion)
- whether hand searching occurred
- was contact made with authors
- languages searched
- exclusion criteria
- whether the search included unpublished material
- List of citations located and included
- Excluded studies and justification for exclusion
- How abstracts and unpublished studies were handled

Searching for studies of risk

Searching for observational studies simply relies on you using terms inherent to your methodology of interest, for example cohort and case-control.

An example of what your search strategy might look like in Medline (OVID platform) if you were searching specifically for cohort and case-control studies would be:

1. exp cohort studies/
2. cohort$.tw.
3. risk.pt.
4. epidemiologic methods/
5. limit 4 to yr=1966-2011
6. exp case-control studies/
7. (case$ and control$).tw.

A list of useful references and pre-defined search filters such as the one here to aid any search of literature employing observational, diagnostic and prognostic studies is available at the InterTASC Information Specialists subgroup homepage at: http://www.york.ac.uk/inst/crd/intertasc/index.htm.

A large study was performed in order to determine the optimal search strategies for clinically sound causation studies in MEDLINE. The study found that to optimize the sensitivity and specificity of a search for studies of causation, the following strategy was the best: Risk.mp OR Mortality.mp or Cohort.tw. The strategy with the best sensitivity was: Risk:.mp. OR Exp Cohort Studies or Between Groups:.tw. The strategy with the best specificity was: Relative risk.tw. OR Risks.tw. OR Cohort stud:.mp (Wilczynski and Haynes, 2003). For reviewers who

wish to be thorough, and have enough time to conduct a large search, using the most sensitive search strategy is recommended. However, clinicians or time-poor researchers may best be suited using a more specific search strategy.

Another useful strategy when searching for studies is to check the articles that were cited by the authors of the study identified (backward citation tracking), or that have cited the study identified (forward citation tracking), through databases that use citation tracking. This has been found to be particularly useful when searching for observational studies (Kuper et al., 2006).

Search terms and databases that may be helpful in constructing a search strategy

Search filters are strategies that identify articles based on criteria such as specified words in the title, abstract and keywords. They can be of use to restrict the number of articles identified by a search from the vast amounts of literature indexed in the major medical databases. Search filters have strengths and weaknesses and may include wildcard characters as described in the example list below:

(i) Strengths: they are easy to implement and can be pre-stored or developed as an interface

(ii) Limitations: database-specific; platform-specific; time-specific; not all empirically tested and therefore not reproducible; assume that articles are appropriately indexed by authors and databases.

Key to terms used in searching in PubMed:

- ab = words in abstract
- exp = before an index term indicates that the term was exploded
- hw = word in subject heading
- mp = free text search for a term
- pt = publication type
- *sh = subject heading
- ti = words in title
- tw = textwords in title/abstract
- ? = in middle of term indicates use of a wildcard
- / = MeSH subject heading (and includes all subheadings being selected)
- $ = truncation symbol
- adj = two terms where they appear adjacent to one another (so adj4, for example, is within four words)

Generic medical/science databases

MEDLINE/PubMED

One of the most widely searched databases is PubMed, but often MEDLINE and PubMed are used interchangeably. There are in fact some important differences. PubMed is updated more quickly than MEDLINE and PubMed indexes more journal titles and includes the database "Old MEDLINE" as well.

MEDLINE (Medical Literature Analysis and Retrieval System Online) is the U.S. National Library of Medicine's main bibliographic database with references to journal articles in biomedicine and the life sciences, containing over 18 million citations. This is the main component of PubMed, which provides access to MEDLINE and some other resources, including articles published in journals that are beyond the scope of MEDLINE, such as general chemistry articles. Approximately 5,200 journals published in the United States and more than 80 other countries are selected and indexed for MEDLINE. A distinctive feature of MEDLINE is that the records are indexed with NLM's controlled vocabulary, the Medical Subject Headings (MeSH).

In addition to MEDLINE citations, PubMED also contains:

- In-process citations which provide a record for an article before it is indexed with MeSH and added to MEDLINE or converted to out-of-scope status.
- Citations that precede the date that a journal was selected for MEDLINE indexing (when supplied electronically by the publisher).
- Some old MEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status.
- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE.
- Some life science journals that submit full text to PubMed Central and may not yet have been recommended for inclusion in MEDLINE although they have undergone a review by NLM, and some physics journals that were part of a prototype PubMed in the early to mid-1990's.
- Citations to author manuscripts of articles published by NIH-funded researchers.

One of the ways users can limit their retrieval to MEDLINE citations in PubMed is by selecting MEDLINE from the Subsets menu on the Limits screen. Other PubMED services include:

- Links to many sites providing full text articles and other related resources
- Clinical queries and Special queries search filters
- Links to other citations or information, such as those to related articles
- Single citation matcher
- The ability to store collections of citations, and save and automatically update searches
- Filters to group search results

Ovid is another search platform that can be used to search the MEDLINE database, as well as many other databases. Table 2 is a brief list of selected features comparing the OVID-Medline platform and the PubMED platform.

CINAHL

CINAHL is a specialist nursing and allied health database from National League for Nursing and the American Nurses' Association. It provides indexing for approximately 3000 journals and includes full text for more than 550 titles, from 1982 onwards and is updated weekly. CINAHL covers nursing, biomedicine, health sciences librarianship, alternative/

Table 2. Features of Ovid and PubMed versions of Medline

Selected Ovid Features	Selected PubMed Features
Common search interface for 11 databases in a variety of convenient groupings.	Access to MEDLINE and PREMEDLINE. Links to NCBI to search Entrez Gene and other genetics databases.
Ability to rerun your search strategy in other Ovid databases.	Searches seamlessly across MEDLINE and PREMEDLINE. Can switch to other NCBI databases via a drop down menu.
Article Linker box connects user to over 30,000 full text online journals available via Health Sciences Library subscriptions. Ovid also provides links to many online full text articles via a "Full Text" link.	Users can switch from "summary" to "abstract" display and click on the Article Linker box to access the Health Sciences Library's online journals. PubMed also provides Links to publisher sites for electronic journals (may require subscription for full-text).
Full text of approx. 270 clinical medical journals.	Users can switch from "summary" to "abstract" and click on the display button to access many of the Health Sciences Library's online journals, denoted by the "Article Linker" box. PubMed also provides Links to publisher sites for electronic journals (may require subscription for full-text).
Can limit to over 15 different specific subject or journal subsets, e.g. AIDS, bioethics, cancer, complementary medicine, dentistry, history of medicine, nursing, toxicology.	Can limit to any of 13 journal subsets.
Use "Find Similar" to automatically retrieve citations on similar topics.	"See Related Articles" creates a search to find articles related to a selected article
Search strategy recovery not available once the user has logged off.	Search strategies are retained in History for eight hours.
Can save searches for subsequent use or may request periodic e-mail updates (Auto Alerts) to a search.	Can register for My NCBI to save searches, set up e-mail updates, and customize filters for displaying results.
Ability to e-mail results to yourself or others.	Ability to e-mail results to yourself or others via the "Send To" e-mail feature
Common limits may be applied from the initial search screen.	Limits link is available on the initial search screen.
Search terms automatically map to MeSH headings.	Search terms map to MeSH headings and are also searched as text words.

Table 2. Features of Ovid and PubMed versions of Medline (*Continued*)

Selected Ovid Features	Selected PubMed Features
MeSH terms are not automatically exploded.	MeSH terms are automatically exploded.
MEDLINE updated weekly; PREMEDLINE updated daily.	PREMEDLINE updated daily.
"Clinical Queries" and "Expert Searches" may be used for quality filtering in MEDLINE and CINAHL.	"Clinical Queries" may be used to retrieve quality research articles. Systematic Reviews and Medical Genetics searches are also available on the "Clinical Queries" page.
"Find Citation" feature can be used to locate a citation when you have incomplete information.	"Citation Matcher" feature can be used to find citations when you have incomplete information.
3 to 32 week time lag from journal publication to Ovid MEDLINE access.	1 to 8 week time lag from journal publication to PubMed access.

complementary medicine, consumer health and 17 allied health disciplines. It is a useful database for primary literature relevant to qualitative reviews. In addition, this database offers access to health care books, nursing dissertations, selected conference proceedings, standards of practice, educational software, audiovisuals and book chapters. It has an internal subject thesaurus with over 7,000 terms 2,000 unique to CINAHL. Although a bibliographic database, the CINAHL database continues to include selected original and full-text material.

EMBASE

EMBASE (Excerpta Medica database, produced by Elsevier Science), is a predominantly European biomedical and pharmaceutical database containing over 19 million indexed records from more than 7000 peer reviewed journals starting from 1974 to date (MEDLINE back to 1966). It adds more than 600,000 items annually with an overlap of titles with MEDLINE of around 60%. It uses its own life science thesaurus called EMTREE. This is a primary source for reliable information from journal articles, reports, conference papers, proceedings, letters, and reviews. Records are scheduled to be online within 10 days of receipt and the database is considered to be a balance to MEDLINE giving more non-USA, particularly European sources. It also contains CAS Registry numbers for chemicals and drugs.

Current Contents

Current Contents is a current awareness database that provides easy access to complete tables of contents, bibliographic information, and abstracts from the most recently published issues of leading scholarly journals. Cover-to-cover, expert indexing provides accurate access to all the information available in journals, not just articles.

Scopus

Scopus is the largest abstract and citation database of research literature and quality web sources. It's designed to find the comprehensive information scientists need quickly and easily. Updated daily, Scopus offers over 16,000 peer-reviewed journals from more than 4,000 publishers, over 1200 Open Access journals, 520 conference proceedings, 650 trade publications, 315 book series, 36 million records, results from 431 million scientific web pages, patent records and "Articles-in-Press" from over 3,000 journals.

Other Specialist databases

Many well known databases such as Cochrane Central Register of Controlled Trials (CENTRAL), PEDro and OTseeker contain specialized content mainly indexing RCTs and therefore when searching for information related to etiology of disease or risk, may not be of much use. There is also likely to be duplication of information contained within larger databases, for example MEDLINE and EMBASE.

PsychINFO

PsycINFO is an abstract database that provides systematic coverage of the psychological literature from the 1800s to the present. (The database also includes records from the 1600s and 1700s.) PsycINFO contains bibliographic citations, abstracts, cited references, and descriptive information to help locate a wide variety of scholarly publications in the behavioral and social sciences.

British Nursing Index:

This UK nursing and midwifery index comes from the partnership of Bournemouth University, Poole Hospital NHS Trust, Salisbury Hospital NHS Trust and the Royal College of Nursing. The database provides references to journal articles from all the major British nursing and midwifery titles and other English language titles.

Academic Search Premier (Ebscohost) Academic Search Premier

Contains indexing and abstracts for more than 8,300 journals, with full text for more than 4,500 of those titles. PDF backfiles to 1975 or further are available for well over one hundred journals, and searchable cited references are provided for more than 1,000 titles. The database contains unmatched full text coverage in biology, chemistry, engineering, physics, psychology, religion & theology and more.

HealthSource: Nursing/Academic Edition (Ebscohost)

This resource provides nearly 550 scholarly full text journals focusing on many medical disciplines. Coverage of nursing and allied health is particularly strong, including full text from *Creative Nursing, Issues in Comprehensive Pediatric Nursing, Issues in Mental Health Nursing, Journal of Advanced Nursing, Journal of Child & Adolescent Psychiatric Nursing, Journal of Clinical Nursing, Journal of Community Health Nursing, Journal of Nursing Management, Nursing Ethics, Nursing Forum, Nursing Inquiry*, and many more. In addition, this database includes the *Lexi-PAL Drug Guide* which covers 1,300 generic drug patient education sheets with more than 4,700 brand names.

Academic OneFile Gale

Is the premier source for peer-reviewed, full-text articles from the world's leading journals and reference sources. *Academic Onefile* has extensive coverage of the physical sciences, technology, medicine, social sciences, the arts, theology, literature and other subjects. With millions of articles available in both .pdf and .html full-text with no restrictions, researchers are able to find accurate information quickly. In addition to all of the traditional services available through *InfoTrac*, Gale offer a number of services offered through collaboration with Scientific/ISI. Mutual subscribers of *Academic OneFile* and *Scientific's Web of Science* and *Journal Citation Reports* are provided access to cited references, digital object identifier (DOI) links, and additional article-level metadata, as well as access to current and historical information on a selected journal's impact factor.

Grey Literature or Deep Web searching

Grey or gray literature is information or documentation that is usually not peer reviewed. It can be in the form of reports, newsletters, academic institution, corporations, blogs, conference proceedings, census reports and non-independent research and as such is a way of obtaining information before it is published. Grey literature may reveal research that has not been indexed within main databases or that has not been through a peer review process. This literature may provide value in some cases; however, it should be used with caution for these very reasons.

Grey literature can be seen as biased because it is often not subject to rigorous review and editing processes. The person either writing or publishing the work may have an undisclosed vested interest in the information. A recognised limitation of grey literature in the context of evidence-based practice is the lack of reliable authenticity, accuracy and validity of information. Data has not been peer reviewed and journals referred to are not ISI listed, thus some clinicians do not advocate allowing the use of information from these sites.

In the conduct of a systematic review however, searching the grey literature is a key means of avoiding publication bias. Publication bias refers to the tendency of journal editors and sometimes researchers to publish studies finding a positive effect rather than research that found no effect. By not searching the grey literature, reviews may be open to missing studies that did not find an effect, and hence, bias the findings of the review. This, along with the fact that it is possible to find rare information found no-where else in electronic form, can make mining grey literature sites a valuable exercise. The search strategies used for grey literature sources should also be documented.

Librarians try to adopt pro-active approaches and are invaluable to aid in finding this material, though web-based searching, self-archiving and open access are helping to facilitate access. With millions of resources available on the Internet, it is difficult to find relevant and appropriate material even with good search skills and use of advanced search engines. Issues of trust, quality, and search skills are very real and significant concerns - particularly in a learning context. Academics, teachers, students and researchers are faced with a complex environment, with different routes into numerous different resources, different user interfaces, search mechanisms and authentication processes.

Search interfaces that may help find grey literature include:

Mednar (http://mednar.com/mednar/) is a one-stop federated search engine designed for professional medical researchers to quickly access information from a multitude of credible sources. Researchers can take advantage of Mednar's many tools to narrow their searches, drill down into topics, de-duplicates, ranks and clusters results as well as allowing discovery of new information sources. Mednar comprehensively searches multiple databases in real time. It utilizes the native search tools available at each of 47 related sites/databases.

WorldWideScience.org http://worldwidescience.org/index.html is another Deep Web search mechanism. *WorldWideScience.org* accelerates scientific discovery and progress by providing one-stop searching of global science sources including European (OpenSIGLE), Asian, Indian, African sources and the database interface has only been in existence since June 2007.

It should be remembered that access to bibliographic databases may depend on subscriptions and the search interface may also vary depending on the database vendor (for example Ovid, EBSCO, ProQuest, etc) or whether you access MEDLINE via the free PubMed interface. The following search engines are very useful for finding health-based scientific literature:

- http://www.scirus.com
- http://www.metacrawler.com
- http://www.disref.com.au/
- http://www.hon.ch/Medhunt/Medhunt.html
- http://www.medworld_stanford.edu/medbot/
- http://sumsearch.uthscsa.edu/cgi-bin/SUMSearch.exe/
- http://www.mdchoice.com/index.asp
- http://www.science.gov/
- http://www.eHealthcareBot.com/
- http://medworld.stanford.edu/medbot/
- http://omnimedicalsearch.com/
- http://www.ingentaconnect.com/
- http://www.medical-zone.com/

There are numerous health information gateways or portals on the Internet containing links to well organised websites containing primary research documents, clinical guidelines, other sources and further links. For example:

- World Health Organisation, http://www.who.int/library/
- National Institute on Alcohol Abuse and Alcoholism, http://www.niaaa.nih.gov/
- Canadian Health Network, http://www.canadian-health-network.ca/customtools/homee.html
- Health Insite, http://www.healthinsite.gov.au/
- MedlinePlus, http://www.nlm.nih.gov/medlineplus
- National Guidelines Clearinghouse, http://www.guideline.gov/index.asp

- National Electronic Library for Health (UK), http://www.nelh.nhs.uk/
- Partners in Information Access for the Public Health Workforce, http://phpartners.org/guide.html
- Clinical guidelines sites

Universities, colleges, institutes, collaborative research centres (CRCs) provide a range of relevant resources and web links already listed. For example, theses or dissertations are generally included on universities' library pages because these are often catalogued by library technicians. Sources of digital thesis records include:

ProQuest Dissertations & Theses Database (PQDT)

With more than 2.3 million entries, the ProQuest Dissertations & Theses (PQDT) database is the most comprehensive collection of dissertations and theses in the world. Graduate students customarily consult the database to make sure their proposed thesis or dissertation topics have not already been addressed. Students, faculty, and other researchers search it for titles related to their scholarly interests. Of the millions of graduate works listed, they offer over 1.9 million in full text format. PQDT is a subscription database, so consult your library for availability.

Dissertation Abstracts Online (DIALOG)

This is a definitive subject, title, and author guide to virtually every American dissertation accepted at an accredited institution since 1861. Selected Masters theses have been included since 1962. In addition, since 1988, the database includes citations for dissertations from 50 British universities that have been collected by and filmed at *The British Document Supply Centre*. *Worldwide Dissertations* (formerly European Dissertations), have been included in the file. Abstracts are included for doctoral records from July 1980 to the present.

Individual, degree-granting institutions submit copies of dissertations and theses completed to University Microfilms International (UMI). Citations for these dissertations are included in the database and in University Microfilms International print publications: *Dissertation Abstracts International* (DAI), *American Doctoral Dissertations* (ADD), *Comprehensive Dissertation Index* (CDI), and *Masters Abstracts International* (MAI). A list of cooperating institutions can be found in the preface to any volume of *Comprehensive Dissertation Index*, *Dissertation Abstracts International*, or *Masters Abstracts International*.

Others include:

- Dissertation Abstracts
- Theses Canada Portal
- Networked Digital Library of Theses and Dissertations (NDLTD)
- Index to Theses

Academic libraries' Online Public Access Catalogues (OPACS) are excellent sources of grey literature in that these catalogues provide access to local and regional materials, are sources for bibliographic verification, they index dissertations, government and technical reports, particularly if the authors are affiliated with the parent organisation or agency as scholars or researchers. Authors, if in academic positions, sometimes have their own web pages.

Researchers working in a specific topic area often already have extensive reference lists that they are prepared to share or names of others working in the same/related fields, for example authors of Cochrane Collaboration or Joanna Briggs Institute (JBI) protocols for reviews that are not yet completed. This is especially useful for clinicians because they know who works in their specific area of interest. Conference series in the area of interest are also useful sources and can be accessed in academic or national libraries due to the legal deposit rule. Many national libraries collect grey literature created in their countries under legal deposit requirements. Their catalogues are usually available on the Internet. Some also contain holdings of other libraries of that country, as in the Australian National Library's Libraries Australia: http://librariesaustralia.nla.gov.au/apps/kss. WORLDCAT is an impressive service that aims to link the catalogues of all major libraries under one umbrella. http://www.worldcat.org/

Mainstream media often reports recent medical or clinical trials and newspaper sites on the Internet may report who conducted a study, where, when, the methodology used, and the nature of the participants to assist in locating an original source.

Other useful tactics include:

- Setting up 'auto alerts' if possible on key databases to learn about new relevant material as it becomes available.
- Joining a relevant web discussion group/list and post questions and areas of interest; contacts may identify leads to follow up.
- Grey literature is increasingly referenced in journal articles, so reference lists should be checked via hand-searching. Hand searching is recommended for systematic reviews because of the hazards associated with missed studies. Hand searching is also a method of finding recent publications not yet indexed by or cited by other researchers.

Finding grey literature on government websites

Generally, most health-related government-sponsored or maintained websites will go to the trouble of showing:

a) how or if their documents are organised alphabetically, topically or thematically;

b) how individual documents are structured, i.e. contents pages, text, executive summary, etc.;

c) database-type search strategies to find them;

d) links to other web sites or other documents that are related to the documents that they produce;

e) when their collection of grey literature has been updated; and

f) documents in PDF or Microsoft Word downloadable form.

A brief case study for searching

Consider a search on the topic: "the risk of heart disease in people with drug & alcohol dependence". With this query you may wish to determine the risk of heart disease in the management of drug and alcohol dependence. The goal of this study is to uncover as many observational studies as possible, and to present a systematic review with meta-analysis.

Step One – Mainstream Database Search

Do your initial research in the mainstream databases, such as:

- PubMed
- EMBASE
- CINAHL
- PsycINFO

There may be a fair bit of duplication between some of these but you should also note down (perhaps as two separate columns) two things: (a) the keywords or terms used, not forgetting to check if the database uses a thesaurus or controlled vocabulary of indexing terms; and (b) the names of institutions, organisations, agencies, research groups mentioned.

The terminology that could be used in various combinations when searching, (including wild-cards and truncation, which may vary from database to database and should therefore be checked), may include the following:

heart disease, cardiac disease, coronary artery disease, chronic heart disease, coronary disease, heart failure, cardiovascular disease, drug, polydrug, substance, alcohol, tran-quilize, tranquilizer, narcotic, opiate, solvent, inhalant, street drug, prescri*, non-prescri*, nonprescri*, abuse, use, usin*, misus*, utliz*, utilis*, depend, addict, illegal, illicit, intox*, risk, cohort, case-control*, observation*. (Note - in the example, the * has been used to indicate either a wildcard or truncation symbol).

Step Two - Finding and Searching Specialised Databases for Grey Literature

Do a Yahoo or Google Search using keywords heart disease, coronary disease, heart or-ganisations, in combination with the terms from your initial database search. Remember that Google.com 'Advanced Search' is best for this part of the search as it allows you to 'limit' your inquiry in many ways (go to http://www.google.com.au/advanced_search?hl=en).

For our topic, here are a few organisations that are relevant to your search:

- ETOH - Alcohol and Alcohol Problems Science Database, referred to as ETOH, http://etoh.niaaa.nih.gov/Databases.htm
- National Institute on Alcohol Abuse and Alcoholism (NIAAA), http://www.niaaa.nih.gov/
- National Institute on Drug Abuse (NIDA), http://www.nida.nih.gov/
- Canadian Centre on Substance Abuse (CCSA), http://www.ccsa.ca/CCSA/EN/TopNav/Home/
- National Cardiovascular Disease Database http://www.acrm.org.my/ncvd/
- American heart Association http://www.heart.org/HEARTORG/

Step Three – Contacting Directories and Organisations

Contacting relevant organisations noted in your mainstream database search is a good way to assess what resources exist in the form of special databases, library catalogues and the like. Some websites have resources providing a 'jumping-off' point for your search deeper into the World Wide Web. Finding the web sites in Step Two and 'digging deeper' into them will enable you to discover the documents they have, and their links to more precise sites with

databases that specialise in drug and alcohol related issues. Examples of these, including from organisations listed in Step 2, are as follows:

- HTA Database, http://144.32.150.197/scripts/WEBC.EXE/NHSCRD/start
- Drug Database (Alcohol and other Drugs Council of Australia), http://203.48.73.10/liberty3/gateway/gateway.exe?application=Liberty3&displayform=opac/main
- Canadian Centre for Substance Abuse, http://www.ccsa.ca/CCSA/EN/Addiction_Databases/LibraryCollectionForm.htm
- Combined Health Information Database (CHID), http://chid.nih.gov/search/

Subject Heading/Keyword-Related Strategies

The following terms/terminology listed below should be considered (but also brainstorm from these to find similar natural language terms and synonyms) for all the other databases that describe evidence of risk. In particular, it is recommended that the terms listed below, derived from MEDLINE be applied to all the databases not already included in the search filters.

MeSH

The following are examples of subject headings (in bold) for evidence of risk and should be used by searching these terms as 'MeSH Headings' (NLM, 2010).

Randomized Controlled Trial – Clinical trials that involve at least one test treatment and one control treatment, concurrent enrolment and follow-up of the test- and control-treated groups, and in which the treatments to be administered are selected by a random process, such as the use of a random-numbers table.

Controlled Clinical Trials – Clinical trials involving one or more test treatments, at least one control treatment, specified outcome measures for evaluating the studied intervention, and a bias-free method for assigning patients to the test treatment. The treatment may be drugs, devices, or procedures studied for diagnostic, therapeutic, or prophylactic effectiveness. Control measures include placebos, active medicines, no-treatment, dosage forms and regimens, historical comparisons, etc.

Cohort Studies - Studies in which subsets of a defined population are identified. These groups may or may not be exposed to factors hypothesized to influence the probability of the occurrence of a particular disease or other outcome. Cohorts are defined populations which, as a whole, are followed in an attempt to determine distinguishing subgroup characteristics.

Case-Control Studies - Studies that start with the identification of persons with a disease of interest and a control (comparison, referent) group without the disease. The relationship of an attribute to the disease is examined by comparing diseased and non-diseased persons with regard to the frequency or levels of the attribute in each group.

Case reports - Clinical presentations that may be followed by evaluative studies that eventually lead to a diagnosis.

Risk - The probability that an event will occur. It encompasses a variety of measures of the probability of a generally unfavorable outcome.

Harm Reduction - The application of methods designed to reduce the risk of harm associated with certain behaviors without reduction in frequency of those behaviors. The risk-associated behaviors include ongoing and active addictive behaviors.

Conclusion

The search strategy of a systematic review can be considered as akin to the data collection phase of a primary research study, and as such, needs to be transparent, replicable and as rigorous, comprehensive and systematic as possible. The search strategy used should aim to identify all relevant international research within an agreed timeframe. This should include databases that will be searched, and the search terms that will be used. In addition to this, it should also specify what research methods/methodologies will be considered for inclusion in the review. This should be clear in the systematic protocol and once the search is conducted, the details of the strategy used in each database and results should be recorded.

Within systematic reviews the search strategy is described as a three-phase process which begins with the identification of initial key words followed by analysis of the text words contained in the title and abstract, and of the index terms used to describe relevant articles. The second phase is to construct database specific searches for each database included in the protocol. The third stage is to review the reference lists of all studies that are retrieved for appraisal to search for additional studies, for example:

> *The search strategy aims to find both published and unpublished studies. A three-step search strategy will be utilized in each component of this review. An initial limited search of MEDLINE and CINAHL will be undertaken followed by analysis of the text words contained in the title and abstract, and of the index terms used to describe articles. A second search using all identified keywords and index terms will then be undertaken across all included databases. Thirdly, the reference list of all identified reports and articles will be searched for additional studies.*

Reviewers are required to state the databases to be searched, the initial key words that will be used to develop full search strategies and if including unpublished studies what sources will be accessed. There should be a statement about the target study type and the range of studies that will be used if the primary study type is not found.

The search strategy should also describe any limitations to the scope of searching in terms of dates, resources accessed or languages. Each of these may vary depending on the nature of the topic being reviewed or the resources available. Limiting by date may be used where the focus of the review is on a more recent intervention or innovation or where there has been a significant change in practice relevant to the question. Date limiting, however, may exclude seminal early studies in the field and should thus be used with caution; the decision preferably endorsed by topic experts and justified in the protocol. The validity of systematic reviews relies in part to access to an extensive range of electronic databases for literature searching. There is inadequate evidence to suggest a particular number of databases should be included, or even to specify particular databases for inclusion. Thus, literature searching should be based on the principal of inclusiveness, with the widest reasonable range of databases included that are considered appropriate to the focus of the review.

Chapter 6:
Selecting and critically appraising studies

Selecting studies

When the comprehensive search for evidence is complete (or as the search progresses in some cases) reviewers must decide which papers found should be retrieved and then subjected to critical appraisal. This initial process is referred to as 'study selection' and identifies those studies that fulfill the criteria for inclusion in the review as defined in the review protocol. Recall that the inclusion criteria for the review stem directly from the PICO method to develop a review question and guide the search strategy used. The process simply addresses the question of whether or not a particular study should be retrieved.

In practice the process of study selection normally involves two stages - retrieval and inclusion. Firstly, two reviewers (first blinded to each other) screen the initial lists of all of the titles and the abstracts identified during the search of the various information databases and other sources they saved in bibliographic citation manager software. To limit the risk of error, the two assessors first make their retrieval decision blinded to each other and then they compare their selections to make a final decision to retrieve the full text document. The number of citations found during the search may sometimes run into the many thousands if the search strategy was either more sensitive than specific or the scope of the review is large. Therefore, an entire team may be involved in study selection process, dividing the labor amongst the paired members. Sometimes it will be difficult or impossible to determine if the reference matches the inclusion criteria of the review on the basis of the title or abstract alone; in this case it is best to err on the side of caution and the full text of the article should be retrieved for further clarification. A second round of screening can then occur on the full text documents where further detail was necessary to determine the appropriateness of inclusion based on the inclusion criteria stipulated in the protocol. The ability to complete this phase on electronic versions of documents also limits the amount of unnecessary printing! Be sure to maintain a record of how many papers were found, retrieved after screening and excluded after closer review.

Investing more time during study selection in careful consideration of the applicability of studies to the review project is recommended. If not, the risk of missing important and relevant evidence related to the review question which was actually identified by the search strategy used will increase. The entire process should be transparent and clear so that if an independent person were to apply the same inclusion criteria to the same list of citations, they would complete the process with the same result of included studies (necessarily limiting the subsequent search by date of the original search). Normally the large number of citations identified by the search reduces significantly during this phase of the review process. All 'selected' studies are then subjected to critical appraisal to determine methodological quality.

Assessment of methodological quality/critical appraisal

The critical appraisal of identified papers is a defining feature of any systematic review irrespective of the evidence the question lends itself to. Critical appraisal is the step in the

review process that underlies the claim that the evidence of the synthesis is based on the best available evidence, not simply all of the evidence available, and increases the validity of any overall estimate of effect quantified from the included studies. Critical appraisal is sometimes described as quality assessment and is principally focused on establishing the validity and reliability of the study and the other issues related to quality described in Chapter 3. In short, an analysis of the extent to which the study conducted minimizes the potential for bias.

Critical appraisal requires the reader to engage fully with a paper, often to consider factors or items that aren't actually reported on, for example the possibility of a relevant confounding factor that the authors make no mention of and neglected to adjust for. Generally speaking, sound understanding of study methodology is a requirement for undertaking a systematic review and effective appraisal, whereas having little knowledge of the topic and research area in question is often viewed as an advantage to limit biases which may arise if reviewers undertaking appraisal already have their own established ideas about the field of research and, even unwittingly, impart then upon the review process. Having said this, involving a member of the review team who is an 'expert' in the field the question pertains to, can be an advantage to truly engage with the research conducted and appreciate nuances identifiable only via expertise in the field, which may in fact impact on validity.

It is preferable to use a clear, well–defined critical appraisal process. To aid in this critical appraisal commonly involves using an instrument or tool to aid in the evaluation of the quality of a given study. Such tools often take the form of a checklist or scale. Using a checklist helps formalize and systematize the process, providing an audit trail for the assessments made (Shamliyan et al., 2010). To use these tools effectively however, requires each appraiser involved in the process having a good understanding of the research design in question. When considering a RCT for example, if an appraiser is unclear on what true randomization entails, it will difficult for them to answer the questions posed by the appraisal instrument confidently, and may reduce the rigor of the process.

The two most commonly used designs for observational studies which will inform evidence of risk, are cohort studies and case-control studies. As mentioned, in a cohort study the comparison groups are identified according to an exposure of interest, whereas in case-control studies, the study groups are chosen on the basis of their disease or outcome of interest. Critical appraisal of observational studies is often more difficult than that of trials and the items in the checklist will understandably be different due to the differences in the defining features of the study's design. Despite the numbers of these instruments in use, there is a distinct lack of standardized tools that have been employed and validated for use with observational research by more than one systematic review in many cases (Shamliyan et al., 2010). Difficulty in appraisal of these studies also arises because standards for their reporting such as the STROBE statement (von Elm et al., 2007), have been in place for much less time than those for trials. As a result important information which appraisal instruments necessarily drive reviewers towards to establish the validity of the study is unfortunately missing.

Below is an example of a commonly used checklist to appraise comparable cohort and case control studies. This example is from the Joanna Briggs Institute (JBI) and is embedded in

Figure 3: JBI MAStARI appraisal tool for comparable cohort and case-control studies

Design: Comparable Cohort / Case Control Studies

Criteria	Yes	No	Unclear	Not Applicable
1) Is sample representative of patients in the population as a whole?	○	○	○	○
2) Are the patients at a similar point in the course of their condition/illness?	○	○	○	○
3) Has bias been minimised in relation to selection of cases and of controls?	○	○	○	○
4) Are confounding factors identified and strategies to deal with them stated?	○	○	○	○
5) Are outcomes assessed using objective criteria?	○	○	○	○
6) Was follow up carried out over a sufficient time period?	○	○	○	○
7) Were the outcomes of people who withdrew described and included in the analysis?	○	○	○	○
8) Were outcomes measured in a reliable way?	○	○	○	○
9) Was appropriate statistical analysis used?	○	○	○	○

their JBI meta-analysis and statistical assessment and review instrument (JBI MAStARI), an online analytical tool to aid in the systematic review of published quantitative data.[1]

Comparative cohort/case-control

Figure 3 describes the list of questions that needs to be considered when examining Comparative cohort/case-control studies.

It is worth noting that the response to these questions in the checklist is either yes, no, unclear or not applicable. As a general guide, 'yes' is appropriate when there is a clear statement within the paper that boldly answers the question in the positive. Similarly, 'no' is appropriate when there is a clear statement within the paper that boldly answers the question in the negative. Unclear refers to when there is no clear statement within the paper in response to the question, or you feel that the response the authors have provided is ambiguous. Despite these general guidelines it will often depend on the nature of the research that defines 'clearly'. It is recommended that reviewers conduct a blind trial appraisal of a sample of their included to studies to define the exact criteria that should be identifiable in the reported study that will determine a response of yes or no for each of the questions asked. Although sometimes necessary, 'not applicable' should be used with caution and sparingly in agreed circumstances. Most of the questions are asked for good underlying reasons to establish validity and as such should not be omitted.

Below is an example of the questions asked by such an instrument and some further information to be considered when interpreting what the question is asking. It is important whilst undertaking the review proper, that within the review team that the process is piloted, particularly to establish what will mark the definable limits of an answer of 'Yes', 'No' and 'Unclear'. For example for question 1 below, what specifically, should appraisers be looking for to be

---

[1] *Available from The Joanna Briggs Institute website http://www.jbiconnect.org/sumari/mastari/index.cfm*

presented in the published paper to decide that 'Yes' the sample used is representative of patients in the population as a whole. Often these criteria or cut-off may differ depending on the review topic. Regardless, both appraisers, or if necessary for the third adjudicator, that they are approaching the questions with the same, clear, established criteria in mind.

1. Is sample representative of patients in the population as whole?

This question concerns how well defined the individuals in the study are, and whether they are representative of the underlying population. The authors should describe and specify their criteria for establishing that the patient has the disorder of interest and how they selected their patient sample. Several biases related to the assembly of the patients can distort the results of a study. The definition of eligibility for entry to the study, such as the disease definition (case-control or cohort) or the nature of the exposure (in cohort) should be given. If the study subjects have been obtained by some forms of sampling (for example, selecting GPs and taking a sample of their patients) the details of this should be clearly explained. These will indicate whether the sample obtained is likely to be representative of a wider group.

2. Are the patients in a similar point in the course of the condition or illness?

The second issue concerns whether the study patients are all at a similar, well-defined point in the course of their disease. Authors should provide a clear description of the stage of disease at which patients entered the study. For instance, since the duration of illness is often associated with outcome, the investigators should report the duration of illness for the sample patients. Within reason, all or most of the study patients should be at a similar point, such as survivors of a first myocardial infarction or patients newly diagnosed with lung cancer.

3. Has bias been minimised in selection of cases and control?

In case- control study, we select groups on basis of outcome- selecting a case group who have already suffered the outcome and a control group. The cases should truly be cases. The inclusion of individuals who do not in fact have the outcome in question within the case group will tend to dilute the case group and bias the results of the study toward the null value. For example, individuals who are taking regular medication will be more likely to have regular contact with doctors. Thus, newly occurring asymptomatic or mild diseases are more likely to be diagnosed. This could create apparent but spurious relationship between the medication and the new mild disease. The willingness of patients to participate in studies and provide information can also result in bias.

Selecting appropriate controls is one of the major challenges of case-control studies. They are generally selected from the same source as the cases, be that the community, general practice or specialist center. The intention is that they should resemble the cases, except that they do not have the disease being studied. Controls should be selected to be similar in terms of age, sex, social class, and area of residence to the cases.

In cohort studies, follow-up studies investigating exposure/intervention cannot easily be interpreted when data have been collected only on the exposed group. Suppose that the cancer risk of industrial pollution were being assessed. Cancers occur spontaneously, so the true question is whether there is an increase in the frequency of cancer among those exposed.

This can be determined only by comparison with some control group who resemble the study group in all ways except exposure.

4. Are confounding factors identified and strategies to deal with them stated?

The major limitation of cohort studies is that the researchers have no control over the group to be investigated. The study group is selected because they have some disease or have been exposed to some potentially noxious agent. Individuals who develop a disease may differ in many ways from those who are disease free. Equally, those who have been exposed to a noxious agent are likely to differ from those who have not. The process of selecting individuals with some defined characteristics can influence the study outcome. Thus, outcomes need to be reviewed to determine which aspects of individual behavior, such as smoking, drinking, diet and exercise, or characteristics of the disease or its management. Furthermore, because the investigator does not manipulate the independent variable as in an RCT, the investigator must be certain of the timing of the exposure and outcome measure. That is the exposure must have occurred prior to the outcome. Having identified these factors, the paper can be inspected to see whether these factors were investigated and, if so, how they were allowed for in the analysis. Regression and, less frequently, stratification are two analytical strategies that reviewers may encounter in cohort studies for example to account for confounding factors (Normand etc al, 2005).

Likewise, in case- control studies, when a disease is related to some factor, attention should be given to other factors that might be related to outcome. Sometimes it is difficult to guess what the confounding factor might be, but the subject should be raised at some point in the discussion section. When confounding factors are identified in a study they should be taken into account in the analysis. There are standard statistical methods to do this. It can be difficult to tell whether these have been used correctly, but at least they should have been used.

5. Are outcomes assessed using objective criteria?

Investigators must provide a clear and sensible definition of adverse outcomes before the study starts. Outcome events can vary from those that are objective and easily measured (death), to those that require some judgment (myocardial infarction), to those that require considerable judgment and may often be difficult to measure (e.g., disability, quality of life). To minimize bias, the individual determining the outcomes should not know whether the patient had a potential prognostic factor. This is not always possible and, for unequivocal events such as death, may not be necessary. However, blinding is essential for outcomes requiring a great deal of judgment, such as transient ischemic attacks or unstable angina.

6. Was follow up carried out over sufficient period?

In follow up studies patients have many opportunities to disappear. Marriage, death, migration, or admission to a long-stay hospital can all result in patients being lost to follow-up. Whatever reasons for it, those lost to follow up are likely to differ from those who remain in view. The greater the extent of this loss the greater the potential for bias. Thus, the key questions are: how great was the loss? And to what extent, given the circumstances of the study, could this influence the results? These issues will be discussed in question 7.

Furthermore, the length of follow up should be reviewed to clarify whether it was long enough to have a reasonable chance of detecting important events. The minimum length of follow up depends on the events being studied. For example, if the study investigated pain and discomfort after discharge from day care surgery, then a few weeks of follow up would be sufficient. However, if the aim as to detect adverse events associated with a new treatment then much longer follow up would be required. Side effects may take several months to develop. Further, it there was interest in diseases like cancer, which can take many years to occur, then a substantially longer follow up would be needed. The question to be asked is: how long could it be until the events of interest occur? The follow up period should be substantially longer than the expected time.

7. Were the outcomes of people who withdrew described and included in the analysis?

Since the presence of a prognostic factor often precedes the development of an outcome event by a long period, investigators must follow patients for long enough to detect the outcomes of interest. Ideally, investigators will succeed in following all patients but this is often not the case. Patients are not usually unavailable for follow-up for inconsequential reasons. Patients may fail to return because they have suffered exactly those events in which the investigators are interested (e.g., they died or have been institutionalized). Conversely, patients who feel entirely healthy may also be less likely to return for evaluation because they feel so well. Simply put, the greater the number of patients unavailable for follow-up, the less accurate the estimate regarding the risk of the adverse outcome.

Under what circumstances does unavailability for follow-up compromise the validity of a study? It is important that you consider the relation between the proportion of patients who are unavailable and the proportion of patients who have suffered the adverse outcome of interest. The larger the number of patients whose fate is unknown relative to the number who have suffered an event, the greater the threat to the study's validity. For instance, let us assume that 30% of a particularly high-risk group (such as elderly diabetics) have suffered an adverse outcome (such as cardiovascular death) during long-term follow-up. If 5% of the patients have been lost, the true rate of patients who had died may be as high as 35%. Even if this were so, the clinical implications would not change, and the unavailability for follow-up doesn't threaten the validity of the study. However, in a much lower-risk patient sample (otherwise healthy middle-aged men, for instance) the observed event rate may be 1%. In this case, if one assumed that all 5% of the patients unavailable for follow-up had died, the event rate of 6% would have very different implications. If the number of patients unavailable for follow-up potentially jeopardizes the study's validity, you should look for the reasons for patients being unavailable, and compare the important demographic and clinical characteristics of the patients who were unavailable with the patients in whom follow-up was complete. To the extent that the reasons for disappearance are unrelated to outcome and the patients who are unavailable are similar to those for whom information is complete, you may feel reassured. If investigators omit information about reasons for unavailability for follow-up or the characteristics of the patients who are unavailable, the strength of inference from the study results will be weaker. The authors should explain the reason for withdrawal.

8. Were outcomes measured in reliable way?

Another crucial aspect of follow up is the way that the outcome was measured. For single events like death or the diagnosis of cancer, measurement is straightforward because the information will be stated in official records. Physiological or biochemical measures of disease status, such as peak flow for respiratory disease or plasma creatinine for renal disease, also provide objective measures but the techniques of measurement need to be rigorous nonetheless.

When clinical judgment is being used, for example to arrive at a diagnosis, there is an opportunity for error. Equally, when patients are being interviewed, the nature of the questioning, either probing or passing on, could influence the answers obtained. Thus, the question to be asked is could the method of obtaining the outcome data materially affect the result? It is important then that the person taking these decisions is blind to the exposure group, otherwise biases might occur.

9. Was appropriate statistical analysis used?

When appraising the appropriate statistical method, the reviewer should attend to several factors including:

- the statistical test matches the research design
- the statistical test matches the level of measurement
- assumptions of the statistical test were met
- the number of measurements and their relevant dependence or independence
- historical changes in the population over time

Because cohort studies are conducted over time and they may or may not involve a comparison group or repeated measures, the statistical test should account for these design features. Parametric statistical tests are based on certain assumptions such as normal distributions and homogenous variability (when more than one measure is made). Nonparametric tests are generally more robust-for example they may be distribution-free. Often, the appraiser must examine the results and make a judgment about assumptions, as they are frequently not directly addressed in reports. The level of measurement influences the type of statistical test. In some studies of exposure the outcome is a simple discrete, categorical variable like the condition is present or absent. Thus the researcher typically reports frequencies and tests proportions (e.g., Odds ratios or relative risk). If the investigator uses a regression analysis, the choice must match the level of measurement as well (eg, logistic regression versus multiple regression).

If cohort or case-control studies follow people for several years they may need to allow for the ageing or other historical changes in study participants. Most diseases increase in frequency with age so the analysis needs to account for these changes coincident to the exposure. Further, if important factors which could influence the outcome have been identified these should also be incorporated into the analysis such as attrition. It can be difficult for the non-statistician to be sure whether the methods are truly appropriate, but it will be possible to check that some attempts have been made to cope with the complexities which have been identified.

Conclusion

Critical appraisal that determines a reviewer's decision to include or exclude studies retrieved from the search is one feature that sets the systematic review process apart from other types of literature reviews. It insures that unreliable data are not synthesized, damaging the value of the systematic review and creating misleading recommendations. Appraisal requires a background in research methods, a keen attention to the details of reports and a standardized approach.

Chapter 7:
Data Extraction

Once the reviewer has made the decision to include studies of sufficient quality, the next step is to extract the relevant data for the review question. There are two types of data to extract: descriptive details of the study and the sample and the actual outcome results due to the exposure.

Study Details

Data extraction begins with recording of the methodology (such as prospective cohort, nested, case-control etc), identifying the setting and describing the characteristics of the participants. When data extraction of the study background detail is complete, the extraction becomes highly specific to the nature of the data of interest and the question being asked in the review. Data extraction serves the same purpose across study designs – to summarise the findings of many studies into a single document. The process involves transferring findings from the original study using a standard, agreed-upon approach for the specific review. This format is essential to minimise error, to provide an historical record of decisions made about the data for the review, and to become the data set for meta-analysis and/or a narrative summary. Using the JBI-MAStARI instrument[2], there is a series of standardized fields to complete in the initial data extraction phase (Figure 4).

Method

A method section is simply a statement of the research design the investigators used in the original study such as case-control. Include enough detail to be certain of the important details of the design like number of groups, type of comparison groups (eg, historical control, concurrent control, etc). The general data collection details can be included such as timing of data collection, number of measures, etc.

Setting

The setting refers to the important details of where the study took place. For example, in an acute care setting of a large metropolitan hospital, a rural community setting, etc.

Participants

Detail the inclusion and exclusion criteria, sampling design and summarize the sample characteristics like age, gender, cultural background and any other relevant information for the conditions under study.

[2] *Available from the Joanna Briggs Institute website http://www.jbiconnect.org/sumari/mastari/index.cfm*

Figure 4: JBI MASTARI Data Extraction Instrument

Number of Participants

Identify the final sample size and the group sizes. Extract these data carefully based on the totals supplied by the investigator who should have accounted for attrition.

Interventions/Exposure:

Carefully describe the exposure for each group including any relevant detail. This might include the geography, duration, type, frequency, etc of the exposure. Extract enough detail that others will be able to understand the complete nature of the exposure.

Outcome Measures:

Provide a description/label for each outcome and its scale. For example, in the case of the second-hand smoke and COPD patients, one outcome is dyspnea. The outcome description might be a breathlessness with scale/measure represented as a univariate visual analogue scale for breathlessness on a continuous scale of 0-100 mm. Another option might be a category ratio scale of breathing effort with an interval scale ranging from 0-10.

Results:

Differentiate dichotomous and continuous variable results. Extract data for each outcome that is relevant to the review question. Carefully record the outcome result and the group size for each exposure and control group. Take care to select the appropriate numbers for the group sizes. Essentially, these outcome numbers form the 2×2 table for data analysis in the case of dichotomous variables. If the outcomes are measure at a ratio level, extract the mean, standard deviation and the group size. If only a standard error (SE) is reported in the original

study, convert the SE to a standard deviation (SD) by multiplying the SE by the square root of the sample size.

Authors Conclusions

This is a simple record of the author's conclusion related to the review question. These conclusions should include a concise claim of the bottom line effect of the exposure.

Reviewers Conclusions

This is the notes of the reviewer's conclusions that may or may not match the original investigator.

Conclusion

The data extraction process of systematic review is similar to data collection in primary research. It needs to be done carefully, reliably and with a standardized approach. The extraction elements include the details of the study bibliographic details, the sample, research design, exposure, outcomes description, results and the author and reviewer conclusions. Extraction should be conducted by two reviewers, checking the reliability of extraction through dual entry.

Chapter 8:
Meta analytical techniques for data synthesis

It is important to note that for a number of reasons explored earlier in this volume related to heterogeneity of data (Chapter 3), often in a systematic review addressing a question related the risk associated to an exposure, meta-analysis will often be impossible (or deemed undesirable) to perform and the review will present the synthesis of the evidence available as a narrative summary with no meta-analysis included.

Data describing risk of an outcome due to an exposure or intervention is almost invariably dichotomous or binary in nature (or "yes-no" outcomes). Most reviewers will be aware of and comfortable with meta-analysis of raw data, often presented in a 2×2 table and accompanying methods such as Mantael-Hanzel and Peto's method for meta-analysis and combination of such data (Blettner et al., 1999, Egger et al., 1997). Often evidence related to the risk attributable to an exposure or intervention however is presented as the summary risk estimate, for example the relative risk, risk ratio and odds ratio are common numerical measures. Whether one or the other is presented will often depend on whether a prospective or retrospective study design was used. The hazard ratio is similarly used to present dichotomous survival data that incorporates time in the outcome measure. Commonly, these are the only measures that will be reported by the primary studies and the reviewer will be left with these values as the only data available for abstraction (Greenland, 1987).

Irrespective of meta-analysis in a systematic review, some reviews that have not statistically combined studies will present these various measures of risk reported by individual studies all in one figure encompassing 'risk' as shown in figure 5 without the overall summary estimate calculated (Lawlor et al., 2003). Such presentation in a review, coupled with narrative synthesis of the results of the primary research, is invaluable to rapidly portray a clear picture of the status of the evidence and a snapshot of the evidence.

Meta-analysis of risk estimates

Despite the impediment to meta-analysis heterogeneity of the published data presents, be it for methodological, clinical or statistical reasons, meta-analysis is almost always possible and can offer a valid means to explore heterogeneity and its impact within a data set (Ioannidis et al., 2008). A combined analysis of individual studies, beyond the outright aim of increased precision due to increased sample size, may be desirable as it allows the exploration of potential confounders and interactions and other modifying effects that may explain the heterogeneity among the included studies.

Sander Greenland (1987) has developed well accepted techniques of how to accommodate meta-analysis when faced with such reporting of solely risk estimates in the published literature. In short, all measures of risk, irrespective of how they are reported, be it odds ratios, relative risks, hazard ratios and the like are all treated the same - simply as measures of risk in general. This commonly asserted assumption in meta-analyses is based on the fact that the

Figure 5: Both crude and adjusted effect estimates from studies are presented. Sub-groups are presented on the basis of study design. The figure incorporates point estimates and 95% confidence intervals (CI) of relative risks, hazard ratios, standardized morbidity ratios and odds ratios. No overall effect estimate (meta-analysis) is calculated. Reproduced from Lawlor et al., 2003 with kind permission[3].
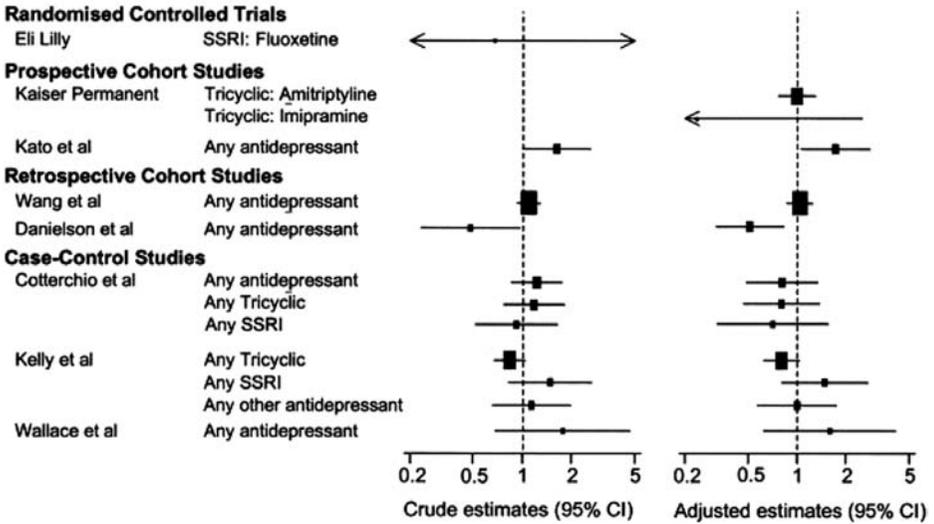
events these studies report on are rare (Greenland, 1987). Generally, where both crude and adjusted estimates of risk are presented in studies, the adjusted estimates are more commonly used (Bekkering et al., 2008). Any differences in adjustments used between individual studies should be discussed in the text of the review and may form the basis for consideration and discussion of heterogeneity observed between included studies.

In meta-analysis of data in the form of crude or adjusted risk estimates, the logarithm of the risk estimate, referred to commonly simply as the relative risk (RR) irrespective of the exact risk estimate reported, are used as the data points in the meta-analysis (Greenland, 1987). As with any meta-analysis, the data points require a measure of variance, in this case the standard error of the mean (SE), to allow combination and calculation of an overall measure of risk. To adequately combine diversely reported outcome data from a range of individual studies in meta-analysis however, data transformation is often essential. The level and extent of transformation necessary can range from simple to quite complex depending on the amount of information supplied by the primary studies and if the necessary data has been provided. The SE value required for meta-analysis can readily be calculated from the reported confidence intervals in studies (Greenland, 1987).

In some cases studies will only present p values accompanying the measure of risk (i.e. RR) presented. In this case the SE can also be calculated, however broader assumptions are necessary to interpret such data and such transformations may be undesirable (Greenland, 1987). Often review authors will predetermine to what extent they will transform data to enable meta-analysis and will use the availability of data presented as the basis for study inclusion in the review project. Such stipulation is frequently observed accompanying published meta-analyses of individual studies (Gandini et al., 2005) but is often not applied as a requirement in a systematic review where some studies may be included in a meta-analysis where and if applicable, whilst others simply appear presented in narrative. In the rare case where some of the included studies do not present a calculated summary estimate of risk, the reviewer is able to calculate the crude RR and its confidence intervals from the raw data.

When data are combined directly from adjusted RR estimates as described here the most common statistical approaches to meta-analysis appear those described using Woolf's method and that of DerSimonian-Laird (DerSimonian and Laird, 1986). These statistical methods are commonly associated with the random effects model of meta-analysis. In the random effects model the underlying assumption is that variability in the data arises from variability within study samples (i.e. between the subjects) and from the differences between the studies also. This is the primary distinction between the random and fixed effects models of meta-analysis, the latter of which, commonly used when applied to meta-analysis of RCTs, considers only within study variation rather than the between study variation. As the random effects model effectively incorporates the method for estimating unexplained variation in the analysis, this model is most frequently applied to 'compensate' for the heterogeneity apparent in observational studies (Blettner et al., 1999). Where there is no heterogeneity present, the results of fixed and random effects models will be similar.

Dose response relationships

As mentioned in Chapter 3, frequently primary published studies investigating risk of an exposure will design the study and present the available data at different levels of the exposure, or in different categories to reflect a 'dose-response' relationship between the exposure and outcome variable. Difficulties will naturally arise if different studies have used different exposure categories and have presented this data in a variety of different ways. Irrespective of this, methods are available to combine the results of individual studies presenting such 'trend' data. Dependent on the type of data presented from such a dose response investigation, accepted methods exist to summarize the data to a consistent risk estimate which can then be subsequently used in meta-analysis (Bekkering et al., 2008). The methods used, with examples of their use, are described by Greenland and Longnecker (1992) and Chêne and Thompson (1996).

Conclusion

The statistical combination of data from observational studies is both advantageous and desirable to inform health care practice and policy regarding the risk associated to an exposure or treatment. As with meta-analysis of RCTs commonly used to inform the effectiveness of an intervention or therapy, the aim is to increase the precision of the overall summary estimate of effect.

Reviewers undertaking a meta-analysis, as a part of their systematic reviews, to further inform the risk attributable to an exposure will discover the process is more complex than a meta-analysis of comparative trials for example. Generally, the outcome data is anticipated to be in the dichotomous format, or binary data, as would be expected to summarize any estimate of risk. However, often studies, be they prospective cohort studies or case control studies, present the overall risk estimate and no raw values for the data from the study participants. This presentation of the data does not preclude meta-analysis however, but rather requires the statistical combination of the logarithm of the risk estimate itself, the log RR and its variance (Greenland, 1987). These values can be calculated from the data extracted in many, but not all cases to allow meta-analysis to be performed.

Published research studies will also often report on a dose response relationship between exposure variable and outcome of interest to strengthen any inferences made. This too can cause problems for the statistical combination of the data depending on how it is presented. Methods exist for expressing such dose response data in a valid and consistent form to allow meta-analysis to be performed.

Chapter 9:
Developing a systematic review report

A systematic review report is the final outcome of any type of review. The report mirrors the components of the original protocol, though written in the past tense, but includes the results and discussion of the review process. Most importantly, the report is a highly transparent record of the processes and outcomes.

The reviewer should include the following elements in the report:

- a comprehensive background justifying the conduct of the review
- objectives and question for the review
- inclusion and exclusion criteria for considering studies,
- clear description of the search strategy including all phases and keywords used and a sample structured search used during phase two as an appendix
- critical appraisal methods including the tools used for each included study design in the appendices
- clear description of the methods of data extraction, includes the tool in the appendices
- a description of the method of meta-analysis and finally a display of the results and a discussion of the method limitations, and implications for clinical practice and research
- appendices.

Results

Like primary studies, commonly, reviewers use subheadings in the results section like:

- description of studies
- methodological quality
- review specific headings related to the exposure and/or outcome
- Forest plots figures of any meta-analyses conducted.

The review of results begins with a discussion and flow chart of papers and their destinies at each stage of the review: the raw total of studies found, the number of duplicates, the number retrieved and excluded, the number included and excluded after full text review, the number included for meta-analysis and narrative synthesis. This type of figure provides readers with a quick summary of the review net output from initial search to final synthesis (see Figure 6). In addition, the reviewer should summarize the types of studies found and the overall quality of the research, this too can be presented in a summary table.

The results section must be organised in a meaningful way based on the objectives of the review and the criteria for considering studies. If a meta-analysis was possible, this forms just one part of the results section. An important aspect of the results is a discussion of the patterns that emerge from the studies included in the review. The reviewer needs to elaborate on these patterns in the discussion of the synthesis. Similar to a report of primary research, the goal is to provide a clear, well-organized representation of the results. Reviewers

Figure 6: Flow chart detailing study review process from search through to meta-analysis
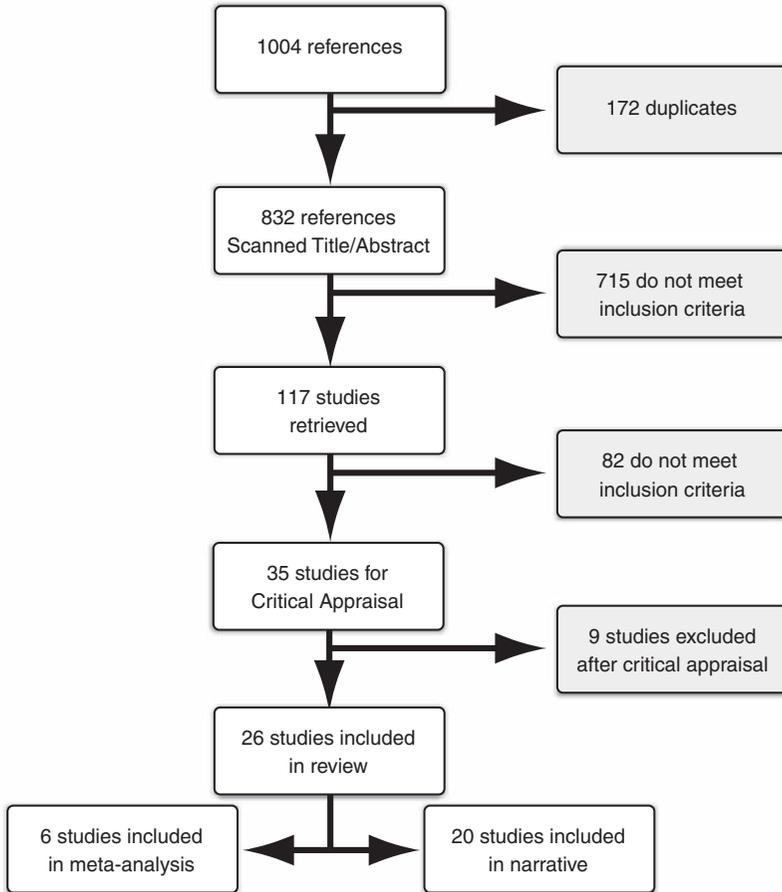
should give particular consideration to the clear descriptions of the types of exposures and outcomes.

Assigning levels of evidence to recommendations

Each study included in the review should be assigned a level of evidence using a standardized hierarchy. This is particularly important if the study designs and the rigor of the studies vary. The evidence table should identify the level of evidence garnered from each individual study. Different organizations have created various levels of evidence based on the ranking or hierarchy of evidence. As an example the Oxford Centre for Evidence Based Medicine levels are shown in Table 3. The first column of the table describes levels of evidence appropriate for evidence of risk or etiology of disease.

Table 3. Oxford Centre for Evidence-based Medicine Levels of Evidence (March 2009)

Level	Therapy/ Prevention, Aetiology/ Harm	Prognosis	Diagnosis	Differential diagnosis/ symptom prevalence study	Economic and decision analyses
1a	SR (with homogeneity*) of RCTs	SR (with homogeneity*) of inception cohort studies; CDR† validated in different populations	SR (with homogeneity*) of Level 1 diagnostic studies; CDR† with 1b studies from different clinical centres	SR (with homogeneity*) of prospective cohort studies	SR (with homogeneity*) of Level 1 economic studies
1b	Individual RCT (with narrow Confidence Interval‡)	Individual inception cohort study with >80% follow-up; CDR† validated in a single population	Validating** cohort study with good††† reference standards; or CDR† tested within one clinical centre	Prospective cohort study with good follow-up****	Analysis based on clinically sensible costs or alternatives; systematic review(s) of the evidence; and including multi-way sensitivity analyses
1c	All or none§	All or none case-series	Absolute SpPins and SnNouts††	All or none case-series	Absolute better-value or worse-value analyses††††
2a	SR (with homogeneity*) of cohort studies	SR (with homogeneity*) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity*) of Level >2 diagnostic studies	SR (with homogeneity*) of 2b and better studies	SR (with homogeneity*) of Level >2 economic studies
2b	Individual cohort study (including low quality RCT; e.g., <80% follow-up)	Retrospective cohort study or follow-up of untreated control patients in an RCT; Derivation of CDR† or validated on split-sample§§§ only	Exploratory** cohort study with good††† reference standards; CDR† after derivation, or validated only on split-sample§§§ or databases	Retrospective cohort study, or poor follow-up	Analysis based on clinically sensible costs or alternatives; limited review(s) of the evidence, or single studies; and including multi-way sensitivity analyses
2c	"Outcomes" Research; Ecological studies	"Outcomes" Research		Ecological studies	Audit or outcomes research

(continued)

Table 3. Oxford Centre for Evidence-based Medicine Levels of Evidence (March 2009)
(*Continued*)

Level	Therapy/ Prevention, Aetiology/ Harm	Prognosis	Diagnosis	Differential diagnosis/ symptom prevalence study	Economic and decision analyses
3a	SR (with homogeneity*) of case-control studies		SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies	SR (with homogeneity*) of 3b and better studies
3b	Individual Case-Control Study		Non-consecutive study; or without consistently applied reference standards	Non-consecutive cohort study, or very limited population	Analysis based on limited alternatives or costs, poor quality estimates of data, but including sensitivity analyses incorporating clinically sensible variations.
4	Case-series (and poor quality cohort and case-control studies§§)	Case-series (and poor quality prognostic cohort studies***)	Case-control study, poor or non-independent reference standard	Case-series or superseded reference standards	Analysis with no sensitivity analysis
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on economic theory or "first principles"

As Table 3 shows, levels of evidence of effectiveness relate to the ability of the research design to control extraneous variables and alternative explanations for outcomes. Studies investigating exposure typically cohort or case-control studies, typically fall within level 2 and 3. As many hierarchies exist it is useful to provide an explanation or table that illustrates the level system utilized in the review. While several hierarchies exist, most place the types of observational designs that are relevant to determining the effect of exposure in the middle

or lower portion of a hierarchy however they vary the ranking of individual study designs. In addition, some hierarchies, like the GRADE group, adjust the ranking based on design based on consideration of the magnitude of the findings and/or flaws in the conduct of the trial.

Discussion and Conclusion

The discussion should include an overview of the results and it should address issues arising from the conduct of the review including limitations and issues discovered in the analysis. No systematic review will centre only on any statistical approaches to synthesis. The reviewer uses the discussion section to survey the patterns that either emerged or did not emerge in the included studies. The reviewer should address any conflicts with conventional wisdom the review brought to light.

The discussion includes a thorough review of the overall quality of the research, the precision of results and examination of the role of biases in the implications of the review. The reviewer uses judgment to determine how the patterns amongst the findings together with the strengths and limitations of the included studies contribute to the conclusions (IOM, 2011).

Conclusions should centre on the implications for practice and for research. These should be detailed and must be based on the documented results, not author opinion. The reviewer should use cautionary language when appropriate. Because of the inherent lack of control over extraneous variables in observational study designs, the reviewer must use caution in making recommendations. Where evidence is of a sufficient level and magnitude of effect, appropriate recommendations should also be made preferably that are clear, concise and unambiguous.

Appendices

Again, as in the initial protocol, the final review report should include references and appendices that support transparency of the systematic review. The references should be appropriate in content and volume and include background references and studies from the initial search. The appendices should include:

- Critical appraisal form(s)
- Data extraction form(s)
- Table of included studies
- Table of excluded studies with justification for exclusion

These checklists should reflect the types of studies, settings, participants, interventions, and outcomes for the review question posed. If systematic review reports are of a high enough standard they may be utilised as evidence upon which to base clinical practice guidelines and policy documents. The table of included studies lists the bibliographic reference and brief reason for inclusion. Similarly, the excluded study appendix lists the bibliographic information and a reason for exclusion. All studies retrieved for appraisal should be detailed in these appendices.

Conflict of interest

A statement should be included in every review that either declares the absence of any conflict of interest, or describes a specified or potential conflict of interest. Conflict of interest

statements should adhere to the guidelines of the International Committee of Medical Journal Editors (ICMJE) for individual authors and project support (http://www.icmje.org/ ethical_4conflicts.html). Additionally, the Committee on Publication Ethics (COPE) have extensive guidelines for conflict of interest statements that are intended to protect the authors as well as the readers, and review authors should ensure they are familiar with and adhere to the principals described within the COPE framework (http://www.publicationethics.org/).

Implications for practice

Implications for practice should be detailed and based on the synthesis of the results rather than opinion. These implications should reflect the cautions revealed in the analysis while clearly and unambiguously stating how the synthesized findings should or should not shape practice. The implications include a statement related to the limitation of the generalizability of the implications.

Implications for research

The reviewer makes statements about the implications for research based on how the review revealed gaps in the primary research or on areas of weakness, such as methodological weaknesses. The reviewer should avoid generalized statements calling for further research. Instead, the implications for research should be specific and linked to the synthesis of results and gaps.

Conclusion

The systematic review report is an important knowledge transfer tool and therefore needs to be written in a manner that is useful to the knowledge user. Plain language abstracts are useful for professionals and consumers alike because they provide a clear, concise summary of the review. Perhaps most importantly, the body of the report is a transparent record of the entire review process. Like primary research reports, the systematic review report is only as valuable as the information provided to the reader. Each step of the science of the systematic review must be meticulously reported in order to live up to the standards of high quality reviews and to allow users to appraise the report for its rigor and utility.

References

Bekkering, G. E., Harris, R. J., Thomas, S., Mayer, A.-M. B., Beynon, R., Ness, A. R., Harbord, R. M., Bain, C., Smith, G. D. & Sterne, J. A. C. 2008. How Much of the Data Published in Observational Studies of the Association between Diet and Prostate or Bladder Cancer Is Usable for Meta-Analysis? *American Journal of Epidemiology,* 167, 1017-1026.

Black, N. 1996. Why we need observational studies to evaluate the effectiveness of health care. *BMJ,* 312, 1215-1218.

Blettner, M., Sauerbrei, W., Schlehofer, B., Scheuchenpflug, T. & Friedenreich, C. 1999. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *International Journal of Epidemiology,* 28, 1-9.

Chêne, G. & Thompson, S. G. 1996. Methods for Summarizing the Risk Associations of Quantitative Variables in Epidemiologic Studies in a Consistent Form. *American Journal of Epidemiology,* 144, 610-621.

Chou, R. & Helfand, M. 2005. Challenges in Systematic Reviews That Assess Treatment Harms. *Annals of Internal Medicine,* 142, 1090-1099.

Clancy, M. 2002. Overview of research designs. *Emergency Medicine Journal,* 19, 546-549.

Dawes, M., Summerskill, W., Glasziou, P., Cartabellotta, A., Martin, J., Hopayian, K., Porzsolt, F., Burls, A. & Osborne, J. 2005. Sicily statement on evidence-based practice. *BMC Medical Education,* 5, 1.

Deeks, J. J., Higgins, J. P. T. & Altman, D. G. 2008. Analysing Data and Undertaking Meta-Analyses. *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons, Ltd.

Dersimonian, R. & Laird, N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7, 177-188.

Doll, R. 2002. Proof of Causality. Deduction form epidemiological observation. *Perspectives in Biology and Medicine* 45, 499-515.

Egger, M., Altman, D. G., Vandenbroucke, J. P. & Group, O. T. S. 2007. Commentary: Strengthening the reporting of observational epidemiology—the STROBE statement. *International Journal of Epidemiology,* 36, 948-950.

Egger, M., Schneider, M. & Smith, G. D. 1998. Meta-analysis Spurious precision? Meta-analysis of observational studies. *BMJ,* 316, 140-144.

Egger, M. & Smith, G. D. 2008. Principles of and Procedures for Systematic Reviews. *Systematic Reviews in Health Care.* BMJ Publishing Group.

Egger, M., Smith, G. D. & Phillips, A. N. 1997. Meta-analysis: Principles and procedures. *BMJ,* 315, 1533-1537.

Egger, M., Smith, G. D. & Schneider, M. 2008. Systematic Reviews of Observational Studies. *Systematic Reviews in Health Care.* BMJ Publishing Group.

Ernster, V. 1994. Nested case-control studies. *Preventative Medicine,* 23, 587-590.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Abeni, D., Boyle, P. & MELCHI, C. F. 2005. Meta-analysis of risk factors for cutaneous melanoma: I. Common and atypical naevi. *European Journal of Cancer,* 41, 28-44.

Green, S., Higgins, J. P. T., Alderson, P., Clarke, M., Mulrow, C. D. & Oxman, A. D. 2008. Introduction. *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons, Ltd.

Greenland, S. 1987. Quantitative methods in the review of epidemiologic literature. *Epidemiologic reviews,* 9, 1-30.

Greenland, S. & Longnecker, M. P. 1992. Methods for Trend Estimation from Summarized Dose-Response Data, with Applications to Meta-Analysis. *American Journal of Epidemiology,* 135, 1301-1309.

Hemingway, H. & Marmot, M. 1999. Psychosocial factors in the aetiology and prognosis of coronary heart disease: systematic review of prospective cohort studies. *BMJ,* 318, 1460-1467.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. 2003. Measuring inconsistency in meta-analyses. *BMJ,* 327, 557-560.

Hoppe, D. J., Schemitsch, E. H., Morshed, S., Tornetta, P. & Bhandari, M. 2009. Hierarchy of Evidence: Where Observational Studies Fit in and Why We Need Them. *Journal of bone and joint surgery,* 91(supplement 3), 2-9.

Ioannidis, J. P. A., Patsopoulos, N. A. & Rothstein, H. R. 2008. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ,* 336, 1413-1415.

Institute of Medicine (IOM) 2011. *Finding What Works in Health Care: Standards for Systematic Reviews,* Washington, DC, The National Academies Press.

Kestenbaum, B. 2009. Cross-Sectional Studies. *Epidemiology and Biostatistics.* Springer New York.

Kheifets, L. I., Afifi, A. A., Buffler, P. A., Zhang, Z. W. & Matkin, C. C. 1997. Occupational Electric and Magnetic Field Exposure and Leukemia: A Meta-Analysis. *Journal of Occupational and Environmental Medicine,* 39, 1074-1091.

Kuper, H., Nicholson, A. & Hemingway, H. 2006. Searching for observational studies: what does citation tracking add to PubMed? A case study in depression and coronary heart disease. *BMC Medical Research Methodology,* 6, 4.

Lawlor, D. A., Jüni, P., Ebrahim, S. & Egger, M. 2003. Systematic review of the epidemiologic and trial evidence of an association between antidepressant medication and breast cancer. *Journal of clinical epidemiology,* 56, 155-163.

Lloyd-Jones, D. M., Larson, M. G., Leip, E. P., Beiser, A., D'agostino, R. B., Kannel, W. B., Murabito, J. M., Vasan, R. S., Benjamin, E. J. & Levy, D. 2002. Lifetime Risk for Developing Congestive Heart Failure. *Circulation,* 106, 3068-3072.

Lloyd-Jones, D. M., Wang, T. J., Leip, E. P., Larson, M. G., Levy, D., Vasan, R. S., D'agostino, R. B., Massaro, J. M., Beiser, A., Wolf, P. A. & Benjamin, E. J. 2004. Lifetime Risk for Development of Atrial Fibrillation. *Circulation,* 110, 1042-1046.

Loke, Y., Price, D., Herxheimer, A. & Group, T. C. A. E. M. 2007. Systematic reviews of adverse effects: framework for a structured approach. *BMC Medical Research Methodology,* 7, 32.

Mamdani, M., Sykora, K., Li, P., Normand, S.-L. T., Streiner, D. L., Austin, P. C., Rochon, P. A. & Anderson, G. M. 2005. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ,* 330, 960-962.

Murabito, J. M., D'agostino, R. B., Silbershatz, H. & Wilson, P. W. F. 1997. Intermittent Claudication : A Risk Profile From The Framingham Heart Study. *Circulation,* 96, 44-49.

National Library of Medicine (NLM). 2010. *Medical Subject Headings* [Online]. Betheseda: US National Library of Medicine. Available: http://www.nlm.nih.gov/mesh/MBrowser.html.

Normand, S.-L. T., Sykora, K., Li, P., Mamdani, M., Rochon, P. A. & Anderson, G. M. 2005. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ,* 330, 1021-1023.

O'Connor, D., Green, S. & Higgins, J. P. T. 2008. Defining the Review Question and Developing Criteria for Including Studies. *Cochrane Handbook for Systematic Reviews of Interventions.* John Wiley & Sons, Ltd.

Owen, N., Healy, G. N., Matthews, C. E. & Dunstan, D. W. 2010. Too Much Sitting: The Population Health Science of Sedentary Behavior. *Exercise and Sport Sciences Reviews,* 38, 105-113.

Pate, R. R., O'neill, J. R. & Lobelo, F. 2008. The Evolving Definition of "Sedentary". *Exercise and Sport Sciences Reviews,* 36, 173-178.

Peters, J. & Mengersen, K. 2008. Selective Reporting of Adjusted Estimates in Observational Epidemiology Studies: Reasons and Implications for Meta-analyses. *Evaluation & the Health Professions,* 31, 370-389.

Renehan, A. G., Tyson, M., Egger, M., Heller, R. F. & Zwahlen, M. 2008. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *The Lancet,* 371, 569-578.

Roberts, J. & Dicenso, A. 1999. Identifying the best research design to fit the question. Part 1: quantitative designs. *Evidence Based Nursing,* 2, 4-6.

Rochon, P. A., Gurwitz, J. H., Sykora, K., Mamdani, M., Streiner, D. L., Garfinkel, S., Normand, S.-L. T. & Geoffrey, M. 2005. Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ,* 330, 895-897.

Sanderson, S., Tatt, I. D. & Higgins, J. P. 2007. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology,* 36, 666-676.

Schardt, C., Adams, M., Owens, T., Keitz, S. & Fontelo, P. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making,* 7, 16.

Schulz, K. F., Altman, D. G. & Moher, D. 2010. The CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and Pharmacotherapeutics,* 1, 100-107.

Shamliyan, T., Kane, R. L. & Dickinson, S. 2010. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of clinical epidemiology,* 63, 1061-1070.

Stephenson, J. M. & Babiker, A. 2000. Overview of study design in clinical epidemiology. *Sexually Transmitted Infections,* 76, 244-247.

Stone, P. W. 2002. Popping the (PICO) question in research and evidence-based practice. *Applied Nursing Research,* 15, 197-198.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., Thacker, S. B. & Group, F. T. M.-A. O. O. S. I. E. 2000. Meta-analysis of Observational Studies in Epidemiology. *JAMA: The Journal of the American Medical Association,* 283, 2008-2012.

Tay, J. & Tinmouth, A. 2007. Observational studies: what is a cohort study? *Transfusion,* 47, 1115-1117.

Tricco, A. C., Tetzlaff, J. & Moher, D. 2011. The art and science of knowledge synthesis. *Journal of clinical epidemiology,* 64, 11-20.

Valentine, J. 2009. Judging the quality of primary research *In:* Cooper, H., Hedges, L. & Valentine, J. (eds.) *The handbook of research synthesis and meta-analysis.* 2nd ed. New York: Russell Sage foundation.

Vandenbroucke, J. P. 2001. In Defense of Case Reports and Case Series. *Annals of Internal Medicine,* 134, 330-334.

Vandenbroucke, J. P., Von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M. & For The, S. I. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med,* 4, e297.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C. & Vandenbroucke, J. P. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Bulletin of the World Health Organization,* 85, 867-872.

Wenban, A. 2001. Epidemiology - The science that cried 'causation' one too many times. *Journal of Vertebral Sublaxation Research,* 4, 75-77.

Wilczynski, N. L. & Haynes, R. B. 2003. Developing Optimal Search Strategies for Detecting Clinically Sound Causation Studies in MEDLINE. *AMIA Annual Symposium Proceedings,* 2003, 719-723.