The contribution of natural experiments to the public health evidence base: Four case studies in evidence synthesis

Michele Hilton Boon

BA, MA, MLIS, MPH

Thesis submitted for the degree of Doctor of Philosophy at the University of Glasgow

MRC/CSO Social and Public Health Sciences Unit

College of Medical, Veterinary and Life Sciences, University of Glasgow

March 2019

Abstract

Background: Natural experiments and related study designs such as regression discontinuity (RD) are of increasing interest to researchers and decision makers because of their potential to address confounding and selection effects better than other observational study designs, with potentially greater generalisability than controlled experiments. Research methods in health have been relatively slow to incorporate natural experiments compared to other fields such as economics and political science, but interest in these methods is growing rapidly.

Objectives: This thesis aimed to (1) investigate the contribution of natural experimental designs to public health research, specifically the evaluation of public health interventions and environmental causes of disease and (2) explore how systematic review methods might be applied to make better use of natural experiments to inform public health and policy.

Methods: The thesis comprises four case studies, including a systematic review of RD studies of health outcomes, a systematic review of RD studies of minimum legal drinking age (MLDA) legislation, development of a critical appraisal tool for RD studies, and a meta-review of endocrine-disrupting chemicals (EDCs) and breast cancer risk. Review protocols were registered in the PROSPERO database.

Results: The first systematic review identified 181 RD studies of health outcomes which spanned a wide range of public health and policy questions, showing that this natural experimental design has been more widely applied than previously appreciated. Thematic analysis of the forcing variables and threshold rules identified patterns of implementation which will aid in future applications of the design. The MLDA review of 17 econometric analyses identified challenges in the synthesis of natural experimental studies. The review identified evidence that MLDA legislation has a causal effect on mortality and on alcohol-related hospital admissions. A ten-item checklist specific to the methodological requirements of RD designs was developed based on standards for RD produced by the What Works Clearinghouse; only 5% of the 181 studies met all ten criteria. The meta-review included 15 systematic reviews of EDCs and breast cancer risk; no primary studies in these reviews were identified as natural experiments.

Conclusions: Natural experiments have the potential to support stronger causal inference through designs that address selection effects and confounding. For these designs to be translated into better evidence to inform decision-making, systematic reviews need to be able to identify and represent in detail the differences among non-randomised study designs. To do this requires further development of systematic review methods in order to synthesise results from econometric models and assess the quality of natural experimental studies.

Table of Contents

List of Tables

List of Figures

List of Accompanying Material

Appendix 1	Protocol: Evaluation of public health interventions using regression discontinuity designs: a systematic review [CRD42015025117]
Appendix 2	Detailed characteristics of included studies for chapter 3
Appendix 3	Detailed critical appraisal results for chapter 4
Appendix 4	Protocol: Endocrine disrupting chemicals and the risk of breast cancer: a systematic review of reviews [CRD42018089344]
Appendix 5	Literature search strategies
Appendix 6	Detailed critical appraisal results for chapter 6

Acknowledgements

I gratefully acknowledge the financial support received from the Medical Research Council for this work under doctoral study grant 1517742. I am also grateful for the support, encouragement, and inspiration that I received from my supervisors, colleagues past and present, family and friends, and my husband. It is a pleasure to take this opportunity to acknowledge all they have done to enable the completion of this work.

My supervisory team of Laurence Moore, Peter Craig, and Hilary Thomson have made my PhD an enjoyable experience and a time of great personal and professional growth. I am grateful to the staff and students of SPHSU who create such a stimulating and positive environment in which to do research. In particular, I would like to thank Frank Popham, Vittal Katikireddi, Mhairi Campbell, and Marcia Gibson for many interesting discussions and their kind encouragement. Mhairi deserves additional acknowledgement and thanks for her work as the second reviewer of the regression discontinuity studies evaluated in chapter 4. Many thanks to Valerie Wells for acting as second reviewer of the systematic reviews appraised in chapter 6.

I would like to acknowledge the clinicians and researchers whose mentoring fostered my interest in evidence synthesis: Neal Maskrey, Jonathan Underhill, Sara Twaddle, Safia Qureshi, and Olivia Wu. At the University of Glasgow, Phil Hanlon and Jacqui Reilly shaped my research interests and passed on their passion for health research and public health.

In 2017 I was the fortunate recipient of a fellowship from the European Respiratory Society that enabled me to undertake further training at the Iberoamerican Cochrane Centre, Barcelona, and the National Institute for Health and Care Excellence, Manchester. I would like to thank my inspiring and supportive mentors at NICE: Judith Thornton, Beth Shaw, and Kay Nolan. I thank everyone at the department of Clinical Epidemiology and Public Health at the Hospital de Sant Pau for the warm welcome and kindness they showed me throughout my stay in Barcelona, particularly Xavier Bonfill, Pablo Alonso Coello, and David Rigau Comas.

Special thanks are due to Eric Medcalf of the University of Glasgow, without whose insight and thoughtful guidance I would not have completed this work.

Finally, I am grateful to my family and friends, especially my husband Stuart, for their encouragement.

Author's Declaration

The research reported in this thesis is my own original work which was developed and carried out in collaboration with others as follows:

The initial proposal for a doctoral project that would use systematic review methods to investigate natural experiments in public health research was put forward by Peter Craig. I conceived the idea to focus reviews on environmental causes of disease and regression discontinuity, then to focus further on breast cancer and MLDA as topics.

In conducting the reviews, I designed and executed the literature searches. Librarians Candida Fenton and Heather Wollege-Andrews advised on the selection of social science and grey literature databases. After performing the initial sift of search results, I selected studies for inclusion with Hilary Thomson acting as second reviewer. I identified the critical appraisal tools to be used and acted as first reviewer for all critical appraisal. Peter Craig and Hilary Thomson acted as second reviewers, which involved screening a 10% sample of studies and appraising 10% of included RD studies. Mhairi Campbell acted as second reviewer for the appraisal of the 17 MLDA studies. Valerie Wells was second reviewer for the appraisal of the 15 breast cancer systematic reviews. The final appraisals presented in the thesis were arrived at following discussion and consensus with the second reviewer.

I have had sole responsibility for the conduct of all other aspects of the research presented within this thesis.

I declare that, except where explicit reference is made to the contribution of others, this dissertation is the result of my own work and has not previously been presented for a higher degree at the University of Glasgow or any other institution.

Signature:

Printed name: Michele Laura Hilton Boon

Definitions/Abbreviations

AMSTAR A MeaSurement Tool to Assess systematic Reviews

DDT dichlorodiphenyltrichloroethane

DDE dichlorodiphenyldichloroethylene

DiD difference in differences

EBM evidence-based medicine

EDC endocrine disrupting chemicals

GRADE Grading of Recommendations Assessment, Development and Evaluation

ITS interrupted time series

MLDA minimum legal drinking age

MRC Medical Research Council

NICE National Institute of Health and Care Excellence

NRS nonrandomised studies

NZ New Zealand

OCP organochlorine pesticide

OR odds ratio

PBDE polybrominated diphenyl ether

PCB polychlorinated biphenyl

PFOA per- and poly-fluoroalkyl substance

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROSPERO: an international database of prospectively registered systematic reviews

RD regression discontinuity

RDD regression discontinuity designs

RoB risk of bias

RS randomised studies

SIGN Scottish Intercollegiate Guidelines Network

TCDD tetrachlorodibenzo-p-dioxin

UK United Kingdom

USA United States of America

WHO World Health Organization

1 Introduction to the thesis

This thesis investigates the potential for natural experiments to be used more widely and more effectively in evidence synthesis in order to better inform public health policy and practice. This investigation uses systematic review methods to determine the contribution of natural experiments to selected areas of the public health evidence base, while also considering how these methods can be applied to ensure that evidence from natural experiments can be recognised, evaluated, and used to inform decision making. The thesis consists of four case studies drawn from three different types of systematic reviews.

The first case study consists of a systematic review of primary studies that use a robust natural experimental design, regression discontinuity (RD), to investigate the health outcomes of interventions or environmental exposures and to consider the applicability of this natural experimental design in public health. This large systematic review is then drawn on and developed to present two further case studies.

The second case study in this thesis focuses on the example of minimum legal drinking age (MLDA) legislation and examines how the results of RD studies can be synthesised and interpreted in the context of systematic reviews of effectiveness.

The third case study considers a specific aspect of systematic review methodology, namely quality assessment, and reports on the development of a critical appraisal checklist for RD studies.

The fourth case study is a meta-review or overview of systematic reviews which considers the environmental causes of disease, specifically the evidence for endocrine disrupting chemicals (EDCs) as a cause of breast cancer. As a case study of the potential use of natural experiments within evidence synthesis, the meta-review examines how previous systematic reviews have identified, evaluated, synthesised, and presented evidence on environmental causes of disease, and with what impact on the strength of evidence and conclusions of the review.

Finally, the discussion chapter considers the implications of the findings for future public health research and evaluation as well as areas for further methodological developments.

Topics for these systematic reviews were selected because they served two purposes. The first was to illustrate the potential for natural experiments to provide robust evidence for questions of importance to public health; the prevention of breast cancer and of alcohol-related harm are of undoubted relevance to policy and practice. These topics were also illustrative of the two types of questions to which natural experimental methods may usefully be applied in public health, namely (1) the evaluation of interventions not amenable to randomisation and (2) the assessment of the environmental causes of disease. At the same time, the topics were chosen to serve a second purpose, namely to demonstrate how systematic review methods can be applied and further developed in order to translate this evidence into a synthesis that represents, in a thorough and balanced way, the findings, strengths, and limitations of natural experiments, given that these studies may come from diverse disciplines and use a variety of methods that are not yet in common use in public health or epidemiology.

Regression discontinuity was chosen as a focus for the first review because it is considered the non-experimental design closest to a randomised trial and therefore has good potential to produce strong evidence of causal effects that should be of interest to decision makers, yet it is also unfamiliar enough within epidemiology and public health research that it is likely to illustrate some of the areas in which existing systematic review methods require development and innovation in order to incorporate evidence from natural experiments.

MLDA was chosen as the focus for the second review on the basis of the protocol for the RD review, which specified that further analyses would be undertaken if a meaningful number of reviews on the same substantive public health topic were identified. MLDA proved to be a fruitful topic for a systematic review of RD studies because the included papers described similar natural experiments with many outcomes in common; furthermore, the studies were reported in sufficient detail to make further synthesis worthwhile and informative in terms of the

challenges of incorporating findings from natural experimental studies into systematic reviews.

Finally, the environmental causes of disease was identified as a focus for the third review on the basis of an Academy of Medical Sciences report (Academy of Medical Sciences and Rutter, 2007). This report recommended natural experimental designs be used to investigate environmental causes of disease and was one of the earliest publications to put forth the argument for greater consideration of natural experiments in the public health evidence base.

1.1 Research question and aims

The overall research question addressed by this thesis is:

- How can natural experiments be incorporated into systematic reviews to provide better evidence for public health policy and practice?

This thesis aims to:

- Identify how RD has been used to investigate research questions of public health relevance
- Investigate the issues RD studies present for a systematic review of the effectiveness of an intervention and how might these be resolved
- Synthesise evidence from RD studies of the effectiveness of MLDA legislation in reducing alcohol-related harms
- Develop methods of assessing the quality of regression discontinuity studies of health outcomes
- Examine how natural experiments have been used in systematic reviews to investigate environmental causes of disease
- Synthesise evidence from systematic reviews of endocrine-disrupting chemicals as a cause of breast cancer

- Apply and further develop systematic review methodology in order to make better use of natural experiments.

1.2 Overview of the thesis

This section briefly outlines the content of the ensuing chapters of the thesis and the appendices.

Chapter 2 defines natural experiments and describes the different study designs that have been used to investigate them. The strengths and limitations of these designs are considered in the context of an account of the development of evidence-based public health. The chapter reviews the literature that has argued for changes to the evidence-based paradigm and greater use of natural experiments in public health research and policy evaluation.

Chapter 3 reports the methods and findings of a systematic review of regression discontinuity studies of health outcomes. This review demonstrates the relevance of this natural experimental design to public health and policy by showing the broad range of topic areas and evaluation questions to which RD has been applied. An analysis of the cut-off rules used for treatment assignment identifies the types of situations in which RD can be used and should facilitate the identification of natural experiments for future research.

Chapter 4 analyses a subset of RD studies from chapter 3 which evaluate minimum legal drinking age (MLDA) legislation as a natural experiment. The chapter presents a synthesis on the protective effects of MLDA laws with regard to mortality and alcohol-related harms including hospital admissions and motor vehicle accidents. The chapter identifies the issues that RD studies present for data extraction, synthesis, and interpretation of findings within a systematic review of health outcomes.

Chapter 5 describes the development of a critical appraisal method for regression discontinuity studies. Existing standards for RD are applied to a sample of studies and adapted into a ten-item checklist which is then applied to the studies identified in chapter 3. The results of appraisal give a comprehensive

picture of the strengths and limitations of this literature and point out the need for improved conduct and reporting of RD studies in health.

Chapter 6 investigates what contribution natural experiments might make to understanding environmental causes of disease by presenting the findings from a meta-review on endocrine-disrupting chemicals (EDCs) as a cause of breast cancer. This meta-review describes how systematic reviews have evaluated and presented evidence from different study designs in reaching their conclusions about EDCs, how the reviews vary in their methods, and how review methods may affect the inclusion and presentation of results from natural experiments.

Chapter 7 discusses the implications of these findings for evidence-based public health. It summarises the findings of the thesis and reflects on the strengths and limitations of the research that has been undertaken. It contains recommendations for additions to the methods of the Cochrane Collaboration, GRADE, and guideline developers such as NICE and SIGN. It offers suggestions for further research and describes future developments of systematic review methodology that would enable those who conduct and use systematic reviews to make better use of natural experiments within evidence syntheses and decision making.

Some details of the methodology and results are supplied in appendices. Review protocols are reproduced in appendices 1 and 4. Appendix 2 provides detailed study characteristics for the 181 RD studies included in the systematic review reported in chapter 3. Appendices 3 and 6 report detailed critical appraisal results. Appendix 5 records literature search strategies.

2 Literature review: Natural experimental methods and public health research

2.1 Chapter overview

This chapter presents a literature review on natural experimental methods and their role in public health research. The chapter begins by providing definitions of public health, evidence-based public health, and natural experiments. It summarises the MRC guidance published in 2012 that raised awareness of the use of natural experiments in evaluating population health interventions. The chapter then describes five methods that can be used to analyse natural experiments: regression discontinuity, instrumental variables, interrupted time series, difference in differences, and propensity score analysis. These methods are of interest because their ability to address selection effects and confounding have the potential to support stronger causal inference than traditional observational methods under certain assumptions. The description of each method is followed by a discussion of its strengths and weaknesses along with examples of application drawn from public health research. The chapter concludes by considering how systematic reviews can further contribute to knowledge of these methods and their use in public health.

2.2 Evidence-based public health

As the thesis addresses the use of natural experiments as evidence for public health decisions, a few definitions are necessary before focussing on natural experiments.

2.2.1 Public health

Public health has been defined as “the art and science of preventing disease, prolonging life and promoting health through the organised efforts of society” (Acheson, 1988). The scope of public health intervention and research therefore encompasses not only disease prevention and health promotion but also the organisation, delivery, evaluation, improvement, and funding of programmes, services, and infrastructure that affect health, together with the policies and legislation that influence and guide these activities. Together with an

understanding of the social determinants of health and the recognition of persistent inequalities in health, public health can be seen as a crossroads or meeting-place of numerous academic disciplines and policy areas. As a discipline, public health has "a long tradition of drawing successfully on other forms of knowledge and insight beyond its own boundaries", which is an asset in dealing with complex problems and emerging threats (Hanlon et al., 2012, p. 9).

2.2.2 Evidence-based public health

Given the broad scope and interdisciplinary character of public health, it follows that the evidence needed to inform public health decisions is likely to be equally wide-ranging. In a public health context, evidence has been defined as "some form of data—including epidemiologic (quantitative) data, results of program or policy evaluations, and qualitative data—to use in making judgments or decisions" (Brownson et al., 2011, p. 6). This definition links evidence (data) to its utility and application, namely in supporting decision-making. Arguably, however, data are of limited utility for decision-making unless they have been analysed and presented in a useable and condensed form, ideally supported by information about their contextual meaning and interpretation, as in a research study or systematic review.

The idea that decisions should be based on "evidence" or empirical research as opposed to anecdote, tradition, habitual practice, or popular belief originated in medicine and subsequently spread to other areas of professional practice and policy (Smith, 2013, p. 42). The originators of evidence-based medicine (EBM) defined it as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996). Jenicek (1997) responded to Sackett's definition of EBM by offering a definition of evidence-based public health (EBPH): "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of communities and populations in the domain of health protection, disease prevention, health maintenance and improvement (health promotion)". Sackett subsequently revised the definition of EBM as "the integration of best research evidence with clinical expertise and patient values", to which Kohatsu et al. responded with a new definition of EBPH: "the process of integrating science-based interventions with community preferences to improve the health of

populations" (2004, p. 419). These definitions all have in common the idea that decision-making benefits from a combination of contextual knowledge relevant to the decision and research evidence that has been assessed for quality.

The importance of taking an evidence-based approach and the perceived poverty of evidence in public health has led to repeated calls to either improve the evidence base or reconsider what may constitute 'best evidence' (Petticrew, 2013). As randomised controlled trials (RCTs) are held to be the highest-quality or 'gold standard' evidence within the hierarchy of evidence-based medicine (Sackett et al., 1996), some have argued that more RCTs ought to be conducted in public health and public policy (Haynes et al., 2012, Macintyre, 2011). The chief benefit of the RCT in terms of supporting decision-making is that random allocation to intervention and comparison groups, performed in a manner not open to bias, is understood to prevent known and unknown confounders from influencing the estimate of the effect of the intervention (Fisher, 1935), thereby producing the evidence most able to support causal inference, that is, the conclusion that the intervention in question independently caused any observed difference in outcome between groups. With evidence from reliable RCTs, decision-makers can be confident that they are choosing to implement and fund interventions that are likely to achieve the desired outcomes.

The obstacles to conducting RCTs in public health are well known and relate to the potential lack of equipoise, feasibility, ethical acceptability, and/or political will (Bonell et al., 2011, Moore and Moore, 2011). Although these barriers are not always insurmountable (Macintyre, 2011, Moore and Moore, 2011), a fundamental problem of public health evaluation and research is how to achieve strong causal inference when a randomised controlled trial is not feasible. Whether the question is one of the causal relationship between an environmental exposure and a disease, or the effectiveness of a policy intervention in changing a particular outcome, non-randomised studies can offer valuable evidence. However, the validity of any non-experimental research is threatened by the potential influence of unobserved factors on the outcome of interest (Academy of Medical Sciences and Rutter, 2007).

Instead of (or in addition to) conducting more RCTs, an alternative approach to improving the public health evidence base is to take advantage of other research

methods and study designs while giving full consideration to their strengths, weaknesses, and reporting quality (Petticrew and Roberts, 2003). Indeed, Sackett et al. (1996) specifically did not restrict the concept of "best evidence" to randomised controlled trials (RCTs) and meta-analyses, or to decisions about interventions, although EBM and its institutions such as Cochrane have become strongly associated with both. Rather, finding the best evidence involves recognising that different study designs provide answers to different types of (clinical) questions, and that the best available evidence may need to be used in the absence of the best possible evidence. In a public health and policy context, natural experimental methods have the potential to incorporate desirable characteristics of the RCT – random allocation, avoidance of selection on observed and unobserved characteristics, control of confounding, and support for causal inference – while overcoming some of the obstacles to conducting RCTs in population health (Petticrew et al., 2005) and providing additional contextual information about real-world implementation and other social or environmental conditions relevant to decision-making.

2.3 Natural experiments: better evidence to inform public health decisions

Definitions of natural experiments vary but have in common the premise that the allocation or delivery of an intervention (such as a policy, programme, or legislative change) is not within the control of the researcher, who instead observes and estimates its effect (Craig et al., 2011, Dunning, 2012, Petticrew et al., 2005). Allocation to the intervention may indeed be random, as when a programme is specifically implemented by a lottery, or it may be held to be 'as good as random' when the researcher can make a strong case that the intervention was randomly taken up by participants (Dunning, 2012). In such situations, the case for causal inference is strong and indeed, Dunning (2012) limits his definition of natural experiments to situations involving random or 'as good as random' allocation; however, this is a narrow definition likely to be realised in only a small number of situations.

A broader definition of a natural experiment is "any event not under the control of a researcher that divides a population into exposed and unexposed groups" (Craig et al., 2017). In this thesis, the term 'natural experiment' refers to such

an event; ‘natural experimental study’ refers to the report of an analysis of such an event by researchers; and ‘natural experimental design’ or ‘natural experimental methods’ refer to the approaches that can be taken in conducting such a study.

2.3.1 Medical Research Council guidance on natural experiments

To support those who conduct, fund, and use such research, the Medical Research Council (MRC) issued guidance on the use of natural experiments to evaluate public health interventions (Craig et al., 2011). The guidance presents seven case studies to demonstrate existing applications of natural experimental methods to a range of public health topics, including child poverty, suicide prevention, alcohol pricing, smoke-free legislation, antenatal care, health service organisation, and early years interventions. The guidance gives priority to building up experience of these methods and notes that systematic review of natural experiments, although demanding, is important to identify and aid in the understanding of promising interventions.

The guidance also provides recommendations for improving the design and analysis, and strengthening causal inference from natural experiments. The guidance recommends three methods – difference in differences, instrumental variables, and regression discontinuity – as representing “a potentially valuable advance” (p. 19) in the analysis of observational data because of their ability to address selection on unobserved variables. These and other key methods used to analyse natural experiments will now be described briefly, including their strengths and limitations, along with examples of their application to public health.

2.3.2 Regression-discontinuity analysis

Regression-discontinuity analysis was first proposed in education research as an alternative to the use of matching to produce a quasi-experimental control group (Thistlethwaite and Campbell, 1960). In this design, subjects are ‘allocated’ to the treatment or control group according to whether or not they meet a threshold value of a ‘forcing variable’ (Imbens and Lemieux, 2008). A forcing variable is a measurement of some attribute whose value can be used to

determine whether or not a subject receives an intervention or exposure. At the defined cut-off value for intervention or exposure, there is a sharp 'discontinuity' in the probability of group allocation. If subjects cannot manipulate the measurement of the forcing variable, and administrators of an intervention cannot manipulate the value of the cut-off, then within a certain range of values or 'bandwidth' of the threshold value, allocation to treatment or control is considered to be essentially random (Dunning, 2012).

The relevance of the RD design to health research has been demonstrated through its application to early versus delayed initiation of antiretroviral therapy in HIV according to CD4 count (Bor et al., 2014) and the prescription of statins according to risk score (O'Keeffe et al., 2014). Indeed, the assignment to treatment according to threshold values or guideline-based rules is common enough in healthcare that regression-discontinuity analysis may be at present underused (Vandenbroucke and le Cessie, 2014). Moscoe, Bor, and Bärnighausen (2015) reviewed the use of regression-discontinuity analysis in epidemiology and public health research. Their search, restricted to a single database (PubMed), identified 18 studies, the majority of which addressed economic or education-related questions.

The chief strength of the regression-discontinuity design is that the element of randomness within the bandwidth on either side of the threshold value supports strong causal inference, at least in theory (Cook and Wong, 2008), negating the effects of both known and unknown confounders (Bor et al., 2014). Additionally, the design can be implemented with routinely collected data, can provide information on the optimisation of clinical treatment thresholds (Bor et al., 2014), and can be used to evaluate policies that impose cut-off values for access, such as age or income level (Imbens and Lemieux, 2008). The limitations of the regression-discontinuity design are that the 'as-if random' quality only applies within the bandwidth close to the threshold, and that performance bias may occur when participants are not blinded to their allocation (Craig et al., 2011).

2.3.3 Instrumental variables

Like regression discontinuity, instrumental variables are used to strengthen causal inference within observational studies, although they are also used in randomised studies, for example to isolate the active ingredient of a complex intervention (Marcus et al., 2012). An instrumental variable or 'instrument' is a variable that meets three criteria: (1) it is correlated with or is a cause of exposure to the independent variable; (2) it does not have a causal influence on the dependent variable, other than through its influence on the independent variable; and (3) it is not correlated with other confounders (Cousens et al., 2011, Dunning, 2012). Economist P.G. Wright is credited with the development of this method in a 1928 analysis of factors affecting supply and demand for flaxseed (Angrist and Krueger, 2001). Instrumental variables are of particular interest in epidemiology because they can be used to adjust for both observed and unobserved confounders (Martens et al., 2006). An example of application to a public health question is a study of the effect of poverty on mental health in Indonesia (Hanandita and Tampubolon, 2014). This study addresses the question of whether the relationship between poverty and increased risk of mental illness is causal or associational by using variability in rainfall as an instrument, on the assumption that rainfall will (in a predominantly agrarian economy) have an effect on poverty but not on mental health. The method is increasingly used in epidemiology; a systematic review identified 90 studies using instrumental variables published between 1994 and 2012 and indexed in either Medline or Embase (Davies et al., 2013).

The idea that an instrumental variable can account for unmeasured or unknown confounders has been described as "an epidemiologist's dream" (Hernán and Robins, 2006) because of the potential to support causal inference from observational data; however, the limitations of the method are not insignificant. In addition to the problem of identifying a suitable instrument and obtaining reliable data for it, the conditions for its use relating to its relationships with other variables and unmeasured confounders cannot ever be entirely empirically verified (Cousens et al., 2011), meaning that the method relies on strong assumptions that cannot be tested from the data (Dunning, 2012). In the view of Hernán and Robins, instrumental variables replace "the unverifiable assumption of no unmeasured confounding…with other unverifiable assumptions" and thus

simply shift the problem of causal inference "to another realm" (2006, p. 364). An articulation of the 'story' or model informing the choice of instrument and the underlying causal theories is therefore necessary (Angrist and Krueger, 2001); however, the previously cited systematic review found poor reporting of the basis for causal inference as well as a number of flaws in the statistical reporting of instrumental variable studies (Davies et al., 2013).

2.3.4 Interrupted time series

A time series is a set of sequential observations or measurements of an outcome taken repeatedly over a period of time. An interrupted time series (ITS) is a type of quasi-experimental design which can be used to analyse a natural experiment in which an event occurs at a defined timepoint and is plausibly expected to have an effect on an outcome, for which time series data are available before and after the event (Lopez Bernal, Cummins, and Gasparrini, 2017). In the simplest ITS design, when the outcome is plotted over time and a segmented regression fitted, a change in the intercept or slope in the post-event period may represent a treatment effect (Shadish, Cook, and Campbell, 2002).

The design is well suited to investigate the effects of policy changes, legislation, and changes to the organisation and delivery of healthcare (such as the introduction of new treatments, quality improvement initiatives, and new models of care). The design was described in a 1968 paper by Donald Campbell and H. Laurence Ross that investigated whether a crackdown on speeding in Connecticut had the effect of reducing road traffic fatalities (Campbell and Ross, 1968). Some recent examples that demonstrate the range of applications in public health include evaluations of the effects of introducing guidelines for antibiotic prophylaxis on the incidence of infective endocarditis (Dayer et al., 2015); the effect of introducing child restraint legislation on child injuries and fatalities in motor vehicle accidents in Chile (Nazif-Munoz, Falconer, and Gong, 2017); the introduction and withdrawal of the Health in Pregnancy grant in England (Adams et al., 2018); and the effect of introducing a surcharge for sugar-sweetened beverages on drinks consumption at leisure centres in Sheffield (Breeze et al., 2018).

The strengths of ITS include the use of administrative and other real-world datasets, with the attendant potential for good external validity, and production of estimates that are not biased by variables that remain constant over time, or that can be adjusted for those variables where data are available and change slowly over time (Lopez Bernal, Cummins, and Gasparrini, 2017). However, the design is also subject to several threats to validity. The most obvious is the possibility that another intervention or change was introduced at the same time as the event of interest and which also affected the outcome, leading to a confounded estimate of the treatment effect. This situation is one specific type of time-varying confounder; another is seasonality or other fluctuations that regularly occur at different times of the day, week, month, or year, such as rush hour traffic or seasonal flu outbreaks, which need to be understood and, if relevant, controlled for in the analysis (Lopez Bernal, Cummins, and Gasparrini, 2017). A further threat to validity is instrumentation if the method of outcome measurement changed during the time period under investigation (Shadish, Cook, and Campbell, 2002).

2.3.5 Difference in differences

The difference in differences (DiD) design involves a comparison of the change in an outcome over time between exposed and unexposed groups following an intervention or change in exposure at a particular point in time (Craig et al., 2012). The effect of the exposure may be estimated additively, as the name suggests, or from a regression that can be adjusted for covariates including time-varying confounders, to which DiD, like ITS, is vulnerable (Angrist and Pischke, 2009).

The design is thought to originate with 19th-century physician John Snow in his classic investigation of the causes of epidemic cholera in London (Angrist and Pischke, 2009, p. 227). Snow was able to show that contaminated water transmitted cholera by comparing death rates between households supplied by two different water companies, one of which moved its water supply further up the Thames and therefore farther from sewage contamination. Importantly, households were not able to choose which company provided their water supply, which had been haphazardly allocated over time (Dunning, 2012), thus ensuring that the analysis of this natural experiment was free from selection effects.

DiD is widely used in econometrics (Imbens and Rubin, 2015) and, with ITS, is one of the most commonly used methods to analyse natural experiments (Craig et al., 2017). Also like ITS, DiD has been used to evaluate a wide range of policy changes, legislation, changes to health systems, and public health interventions. Recent examples include DiD analyses of the effect of achieving Foundation Trust status on hospital performance in England (Verzulli, Jacobs, and Goddard, 2018), the impact of the Affordable Care Act on contraceptive prescriptions (Becker, 2018), and the effect of a school-based public health outreach programme on insurance enrolment and well-child exam uptake (Jenkins, 2018).

2.3.6 Propensity score analysis

A further approach to strengthening the causal inference possible from observational data involves the extension of regression modelling to investigate and adjust for selection on observables in non-random treatment assignment. The methods were developed within econometric structural equation modelling by Heckman in the 1970s, for which he was eventually awarded a Nobel Prize in economics, and within statistics in the 1980s by Rosenbaum and Rubin (Guo and Fraser, 2010, Rosenbaum and Rubin, 1983). These methods are known collectively as propensity score analysis, reflecting the central element of probabilistic modelling of participants' propensity to choose or be allocated to treatment or control groups. In a nonrandomised study, a propensity score can estimate the probability of a participant's group allocation given the values of a set of covariates measured prior to the start of the study (Shadish and Steiner, 2010). This information can then be used in regression modelling of the outcome data to match controls or adjust for selection bias in an attempt to imitate the same balance on pretest covariates that would be achieved through randomisation (Shadish and Steiner, 2010). Rosenbaum and Rubin (1983) argued that such adjustment can produce an unbiased estimate of treatment effects.

Systematic reviews of the use of propensity score methods indicate that the primary application of these methods in the health literature has been in surgical and pharmacological studies (D'Ascenzo et al., 2012, Gayat et al., 2010, Weitzen et al., 2004) – which was also the finding of the systematic review of instrumental variables mentioned previously (Davies et al., 2013). However, instances of application to public health questions also include a natural

experiment in neighbourhood violence reduction in Colombia (Cerda et al., 2012) and evaluations of changes in health care payment policies (Cheng et al., 2012, Stuart et al., 2014).

The chief advantages of propensity score modelling, in addition to improved causal inference, are simplified management of multiple covariates (Guo and Fraser, 2010) and improved matching of treatment and controls (Craig et al., 2011). As these methods are model-based rather than design-based, several caveats apply. The quality of the analysis depends crucially on adequate pretest measures of the covariates influencing the selection process and on sufficient overlap in propensity score values between treatment and control groups (Shadish and Steiner, 2010). Findings are mixed as to whether propensity score analyses produce accurate estimates of effect compared to randomised trials (Kuss et al., 2011, Peikes et al., 2008) and some studies have found that they produce no better estimates than standard regression modelling (Shadish and Steiner, 2010). A further caveat is that propensity score analyses are unable to address unmeasured confounding.

2.4 Investigating the application of natural experiments in public health through systematic reviews

Although natural experiments and related methods have been promoted in the MRC guidance as potentially providing a desirable quality of evidence for public health, questions remain as to why these methods are not more widely used and to what extent they can fulfil the "epidemiologist's dream" of unbiased causal inference from observational data. Systematic review is a method that can be used to determine the characteristics of the use of particular study designs in a given field and to promote new methodologies (Petticrew and Roberts, 2006). The MRC guidance states that systematic reviews of natural experimental studies in public health are important in order to identify interventions that could be further developed, to help with the interpretation of natural experimental evidence, and to synthesise available estimates of effectiveness (Craig et al., 2011, p. 23). The guidance also recognises that such systematic reviews face difficult methodological challenges in needing to deal with multiple study designs, search a wide variety of sources, and address complex risks of bias (p. 23).

This thesis proceeds in chapter 3 to use systematic review as a method to investigate the availability of natural experimental studies in public health by comprehensively identifying and describing RD studies of health outcomes. The thesis then demonstrates how systematic review methods can be used to incorporate RD studies into evidence synthesis (chapter 4) and to interrogate the quality of such studies (chapter 5).

2.5 Chapter summary

This chapter has described evidence-based public health as a context in which natural experimental studies are of interest as potentially providing the 'best evidence' needed to support public health decision-making, particularly when RCTs are not available or not feasible. Public health has been described as broad in scope and interdisciplinary in character, needing evidence from many research areas in order to address the wider determinants of health. Several natural experimental designs have been examined for their strengths, limitations, and application in public health research. The next chapter takes one of these designs which can provide strong support for causal inference, regression discontinuity, and asks how it has been used to provide evidence of relevance to public health decisions, i.e. to investigate the effects of interventions and exposures on health outcomes.

3 Regression discontinuity designs in the evaluation of health interventions, policies, and outcomes: a systematic review

3.1 Chapter overview

This chapter reports a comprehensive systematic review of the application of RD designs in research of the health effects of any interventions or exposures, including social, medical, and public health interventions, environmental exposures, and public policy. Based on a published protocol and a search of 32 databases and grey literature sources, the review identifies 181 studies that apply an RD design in health-related research, more than five times the number of studies identified by a previous review of RD studies that searched only one database (PubMed). A thematic analysis of the underlying natural experimental designs shows that a relatively small number of different types of forcing variables and threshold rules has produced applications across a wide spectrum of research questions relevant to public health and policy, with little replication of the same design to answer similar questions in different settings. Therefore, the review concludes that RD has the potential to be more widely applied in the evaluation of social and public health interventions and policy.

3.2 Aims

This chapter aims to conduct a comprehensive systematic review in order to determine how RD designs have been used in health research. The chapter aims to map the use of RD designs in settings and policy areas relevant to public health by answering the following research questions:

1. How and in what areas of research have RD designs been applied to evaluate the health effects of interventions or exposures?

2. What forcing variables and threshold rules have been used in RD studies of health-related outcomes?

3. What is the quality of reporting in studies using RD designs to evaluate health-related outcomes?

3.3 Background

The regression discontinuity (RD) design was first proposed by Thistlethwaite and Campbell (1960) based on the intuition that, given an eligibility rule based on a cut-off value for a continuous variable whose value cannot be precisely manipulated by participants or administrators, treatment assignment will be effectively random within a certain bandwidth on either side of the cut-off; therefore, the causal effect of the treatment can be estimated by comparing outcomes for groups just above and just below the cut-off, without any bias due to unobservables (Imbens and Lemieux, 2008, Thistlethwaite and Campbell, 1960).

An illustrative example of the implementation of the RD design can be found in a study of the effect of receiving a diagnosis of hypertension on health behaviour (Zhao, Konishi, and Glewwe, 2013). A nationally representative longitudinal survey, The China Health and Nutrition Survey, collected data on individuals' dietary patterns, socioeconomic characteristics, and health status. Trained investigators measured (among other biomarkers) the participants' blood pressure; participants with systolic blood pressure above the diagnostic threshold of 140 mmHg were informed that they had hypertension. Zhao, Konishi, and Glewwe recognised in this situation a natural experiment that could be analysed with an RD design in which systolic blood pressure is the forcing variable. As the authors explain, "Since individuals cannot precisely control their blood pressure, among those with blood pressure readings near the cutoff, some randomly are above it while others randomly fall below it, which can be regarded as a random assignment of hypertension status. Because the consumption patterns and other behaviors are likely to be almost identical for the samples right below and right above the cutoff, the difference in the outcomes between these two groups may be used to estimate the treatment effect – i.e. the effect of being informed that one has hypertension" (p. 368).

The study authors addressed the potential for bias in their study in several ways. First, they checked the assumption that participants could not manipulate their value of the forcing variable, observing that people cannot precisely control their systolic blood pressure. Next, they checked the distribution of the forcing variable in the sample, presenting the data graphically and demonstrating that

there is no evidence of 'heaping' near the cut-off, which might suggest manipulation by the survey investigators. They also checked the distribution of other observed variables in the diagnosed and undiagnosed groups, to see if any systematic differences at baseline might contribute to explaining any difference observed in the outcome of interest. They considered and addressed the potential for attrition bias (loss to follow-up within the original survey). In estimating regressions, they investigated whether the results were sensitive to model specification. Finally, they performed a type of falsification test by checking whether their models detected a false 'treatment' effect at other, non-threshold values of systolic blood pressure (120, 130, etc.); no statistically significant effects were detected at any of the 'placebo' cut-off values. The study concluded that receiving a hypertension diagnosis led to changed dietary behaviour in the form of reduced fat intake.

RD is attractive because it allows the evaluation of causal effects of interventions or exposures using non-experimental data; furthermore, the method requires relatively weak assumptions that can be empirically tested. By using administrative data, existing surveys, or government statistics as well as real-world treatment assignment rules, RD can be implemented efficiently and can avoid the criticism of limited external validity sometimes directed at randomised controlled trials. The main limitation of the design noted in the literature is the need for larger sample sizes than in randomised experiments (Lee and Lemieux, 2010).

Following its initial presentation in educational research in the 1960s, uptake of the design was limited, partly due to a belief that few situations existed in which it could be applied, until its use became common among economists (Cook, 2008). Lee and Lemieux (2010) reviewed the use of RD in the economic literature with the aim of identifying in what topic areas RD had been applied and where cut-off rules could be found. They identified 60 studies, half of which related to education or labour economics, with the remainder spanning diverse topics in political economy, health, crime, the environment, and other subjects. Lee and Lemieux described cut-off rules as emanating from four types of situations: necessary or intentional discretisation (in the allocation of a limited

resource), and nonrandomised discontinuities based on age or geographic boundaries.

Moscoe, Bor, and Barnighausen (2015) argued that RD is likely to be useful in health research because the use of cut-off rules for treatment assignment is common. Their review, based on a search of one database (PubMed), identified 32 studies from medicine, epidemiology, or public health that used an RD design, of which two involved interventions to improve physical health. Accordingly, they argued that RD is likely underutilised in these fields. They evaluated studies on a scale of 0-5 based on the presence of key elements of "good RD practice" and found that a histogram of the assignment variable was the most commonly omitted element.

Most recently, a review in the BMJ (Venkataramani, Bor, and Jena, 2016) presented 13 studies as examples of RD in healthcare. These examples were used to illustrate the application of time, age, programme eligibility, geography, and therapeutic assignment rules in the design and to support an assertion that the design could be applied usefully and widely in clinical medicine and health policy. The review did not use systematic methods.

3.4 Methods

The review protocol was published in the PROSPERO international prospective register of systematic reviews (reference number CRD42015025117). The protocol is reproduced in Appendix 1.

3.4.1 Inclusion criteria

Primary, empirical studies were included from any field of research that (1) used a regression discontinuity design and (2) had an outcome that measured any aspect of physical or mental health or wellbeing.

3.4.2 Search strategy

The search encompassed 32 electronic databases for publications containing the phrase "regression discontinuity" or "regression-discontinuity" in title, abstract, keyword, or (where available as a search option) publication full text. No index

terms were identified that corresponded to RD. The date range covered was 1 January 1960 (year of first publication describing RD methods) to 15 March 2015. No language restrictions were applied. The databases were selected in consultation with expert librarians experienced in systematic review in public health and the social sciences to ensure broad coverage of disciplines relevant to social determinants of health and to public policy, particularly those such as educational research and economics in which RD designs are more commonly used than in health. This approach also reflects previous findings that health databases are not sufficient for comprehensive searches on the health effects of social interventions (Ogilvie et al., 2005). The list of 32 sources searched appears in Table 3.1.

Table 3.1. Databases searched for the systematic review of regression discontinuity studies, by subject area

Health: CINAHL, Cochrane Library, Embase, HMIC, King's Fund Publications, MEDLINE, NICE Evidence Search, POPLINE, PsycINFO, TRIP
Social Sciences: ASSIA, Business Source Premier, EBSCO Professional Development Collection, EconLit, ERIC, International Bibliography of the Social Sciences, Social Care Online, Social Services Abstracts, SocINDEX, Sociological Abstracts
Full Text: Google Scholar, Scopus, Web of Science
Grey Literature: EThOS (British Library Electronic Theses Online Service), Idox Information Service, NTIS, Open Grey, ProQuest Dissertations and Theses, EconPapers (RePeC), US Environmental Protection Agency document repository, WHO Institutional Repository, World Bank Documents and Reports

Reference lists of included studies, review articles, and textbook chapters on regression discontinuity design were hand-searched to identify additional studies.

3.4.3 Study selection

Retrieved references were compiled in an EndNote X7 library and duplicates were manually removed. A random 10% sample (random number sequence generated in Stata version 13) was screened by two reviewers independently for eligibility based on the record title and abstract. Results were compared and disagreements were discussed to clarify the application of the inclusion criteria. After discussion, agreement was 100%. Following this piloting of the study selection process, all references were screened by one author and reasons for exclusion were recorded in EndNote.

3.4.4 Data extraction

The aim of extracting data about publication characteristics was to identify discipline-related patterns in study design, quality, and publication trends. A coding framework was designed to record information extracted from the full text of each included study about the publication, research topic, study design, and outcomes. Each study was given a unique identification number derived from the first author's surname, year of publication, and publication type. Year of publication and publication type were coded as separate fields, with publication type categorised as journal article, working paper, thesis, report, conference paper, or conference abstract. For journal articles, the academic discipline of the journal was additionally described in a method derived from Stuckler et al. (2014), who used the Web of Science category assigned to the journal in order to analyse citation patterns by discipline. In fact, Web of Science typically assigns two or more categories to each journal in its database, without distinguishing a primary classification. Accordingly, all categories assigned to each journal were recorded and then seven groupings were created as follows:

- Health economics: any journal indexed under both an economics category and a health category (including public health)

- Public health: any journal indexed as "Public, Environmental & Occupational Health" and not under an economics category

- Economics: any journal indexed under an economics category, but not a health category

- Psychiatry and psychology: any journal indexed as “Psychiatry” or under any psychology category

- Medical: any journal indexed under any medical specialty (Surgery, Endocrinology, Medicine General and Internal) and none of the above categories

- Other health sciences: any health category not covered by the above (Health Policy and Services, Healthcare Sciences and Services, Nutrition and Dietetics)

- Other social sciences: any category not covered by the above (Political Sciences, Public Administration, Demography, Multidisciplinary Sciences).

The study design was described as sharp or fuzzy RD and any additional designs (such as difference-in-difference or instrumental variable) used in the paper were noted. The country of authorship was recorded as the country of the institution to which the first author belonged at the time of publication. The country of origin of the study data was also recorded. The implementation of RD in the study was described in terms of the forcing variable used, the intervention or exposure under investigation, the health-related outcome(s) measured, whether a primary outcome was specified, and whether a study protocol was reported. Coding was performed by one reviewer.

3.4.5 Quality assessment

The purpose of quality assessment in this review was to describe the strengths and limitations of the literature and thereby enable users, producers, and funders of RD studies in health to understand, recognise, and address quality of conduct and reporting RD. The purpose was not to exclude studies, to identify risk of bias, or to inform meta-analysis of effect estimates. Accordingly, an appraisal tool was sought that was specific to RD and allowed detailed investigation of the methodology of RD studies.

What Works Clearinghouse (WWC) is an online resource centre funded by the United States Department of Education to support reviews of the effectiveness of educational policies and interventions. At the time of conducting this review, the WWC standards were the only publicly available quality assessment tool specific to the design and reporting of regression discontinuity studies (Schochet et al., 2010) (see section 5.4.1). The WWC tool offers detailed criteria and has three additional strengths: it is relatively short and simple to use; it was developed by experts in RD methodology; and it has screening questions that ensure the study correctly employs the RD design.

The WWC tool allows users to determine whether a study meets an overall standard of quality. Accordingly, each of the four standards in the RD tool involves judgments to determine whether the standard has been met, not met, or met with reservations based on whether a combination of various criteria are satisfied. The tool was piloted on 15 studies with two appraisers (MHB and MC) evaluating each study independently. The tool was easy to use, there was little disagreement between appraisers, and the few items of disagreement were easily resolved upon discussion. However, all of the studies 'failed' the overall quality standard because of failing to meet standard 2 (attrition). Both appraisers agreed that most studies based on population or administrative data would 'fail' in this way and that such judgments would not be helpful in meeting our aims of describing quality. Therefore, the WWC tool was adapted by only looking at whether the various criteria were satisfied and not whether the standards were met, not met, or met with reservations. For all included studies, answers of yes/no/unclear were recorded for the three screening questions and yes/no for the seven quality criteria extracted from the tool. These questions and criteria were not used to exclude studies from the analysis. Following the pilot, a 10% sample was appraised by two reviewers and, with satisfactory agreement on interpretation of the criteria obtained following discussion, the remainder of the studies were evaluated by a single reviewer. This process of adapting and developing the critical appraisal method for RD is described in more detail in chapter 5.

3.4.6 Synthesis methods

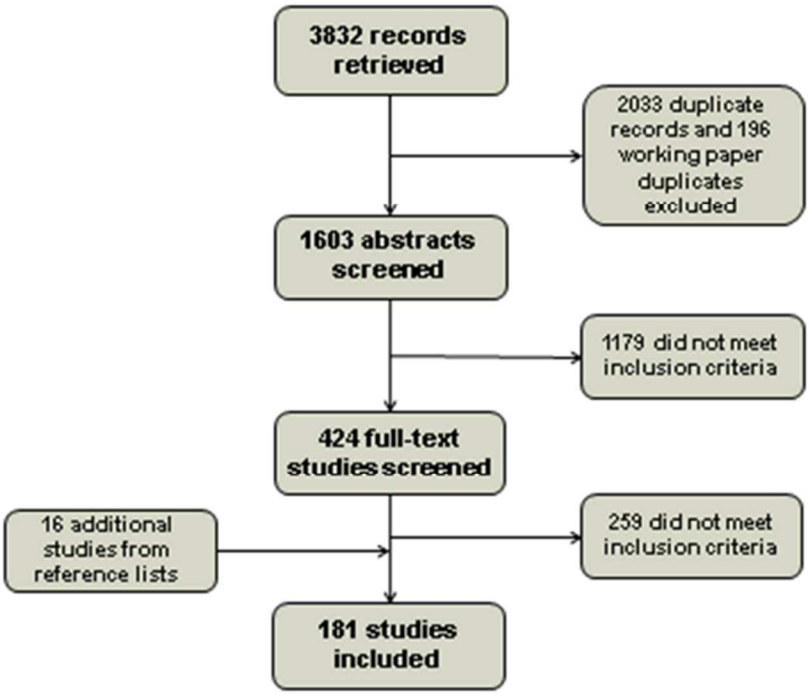
The synthesis methods reflect the aims of the review, namely to map and describe the implementation of RD designs in the investigation of health outcomes. As the review was not designed to identify studies that answer a particular question of effectiveness, no meta-analysis was planned. Instead, a narrative synthesis was undertaken that aimed to identify and describe patterns and commonalities across studies. The main method of narrative synthesis used was thematic summary, in which a descriptive coding framework is developed to allow the grouping of studies in order to compare their characteristics (Gough, Thomas, and Oliver, 2012). Extracted data were presented in tables organised by research topic themes. Additional themes were developed to describe commonalities among the RD designs in terms of forcing variables and cut-off rules. Numbers of publications by year, by topic, and by academic discipline were tallied to enable identification of trends in the use of RD.

3.5 Results

3.5.1 Characteristics of included studies

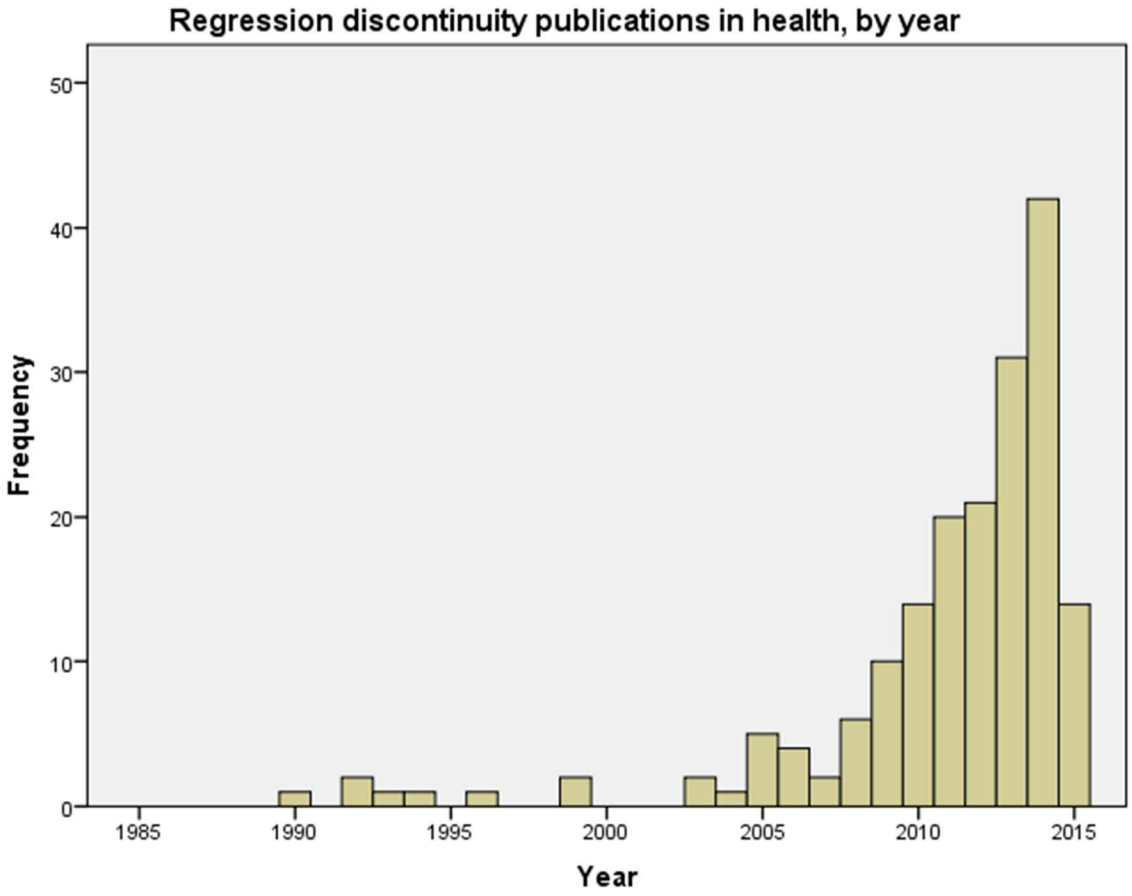
The searches of 32 databases resulted in 3832 records retrieved, of which 2033 were duplicate records and 196 were working paper versions of studies subsequently published as journal articles. The titles and abstracts of the remaining 1603 unique records were examined for evidence of RD design and relevance to health outcomes, of which 1179 did not meet the inclusion criteria. The full text of the remaining 424 studies was obtained and assessed against the inclusion criteria, resulting in the exclusion of a further 259 papers. The reference lists of the included studies were checked for additional references, as were the reference lists of review articles on RD. Sixteen additional studies were identified in this way. Figure 3.1 shows the study selection process as a flowchart. In total, 181 studies were included in this review.

Figure 3.1. PRISMA flowchart for systematic review of regression discontinuity studies



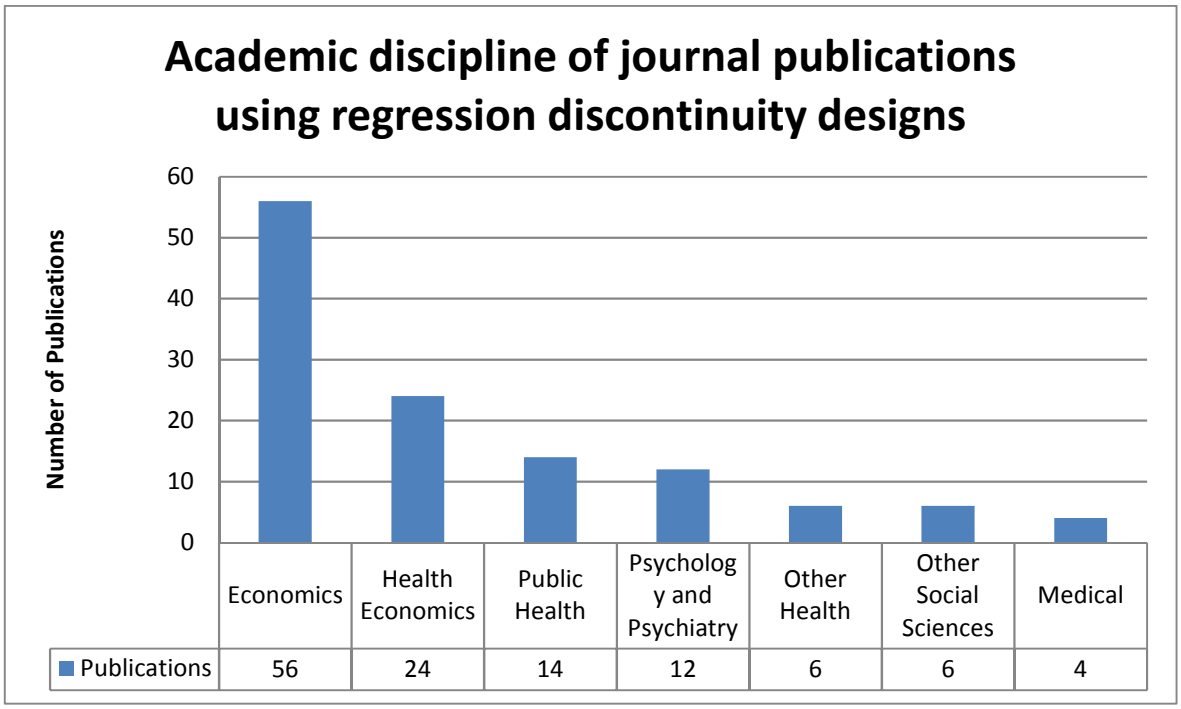
The use of RD designs in health research is increasing over time, with the greatest increase in output taking place in the past five years. The number of studies published by year (Figure 3.2) shows that initially, in the decades following Campbell's first publication describing the design, few studies used RD to investigate health outcomes. The earliest publication relating to health appeared in 1990, with only 28 studies published before 2009. In that year, ten publications appeared and interest increased each year for the subsequent five years, reaching a high of 42 publications in 2014. Results for 2015 are limited to the first quarter only.

Figure 3.2. Histogram of regression discontinuity publications by year
Data for 2015 is limited to January-March only.



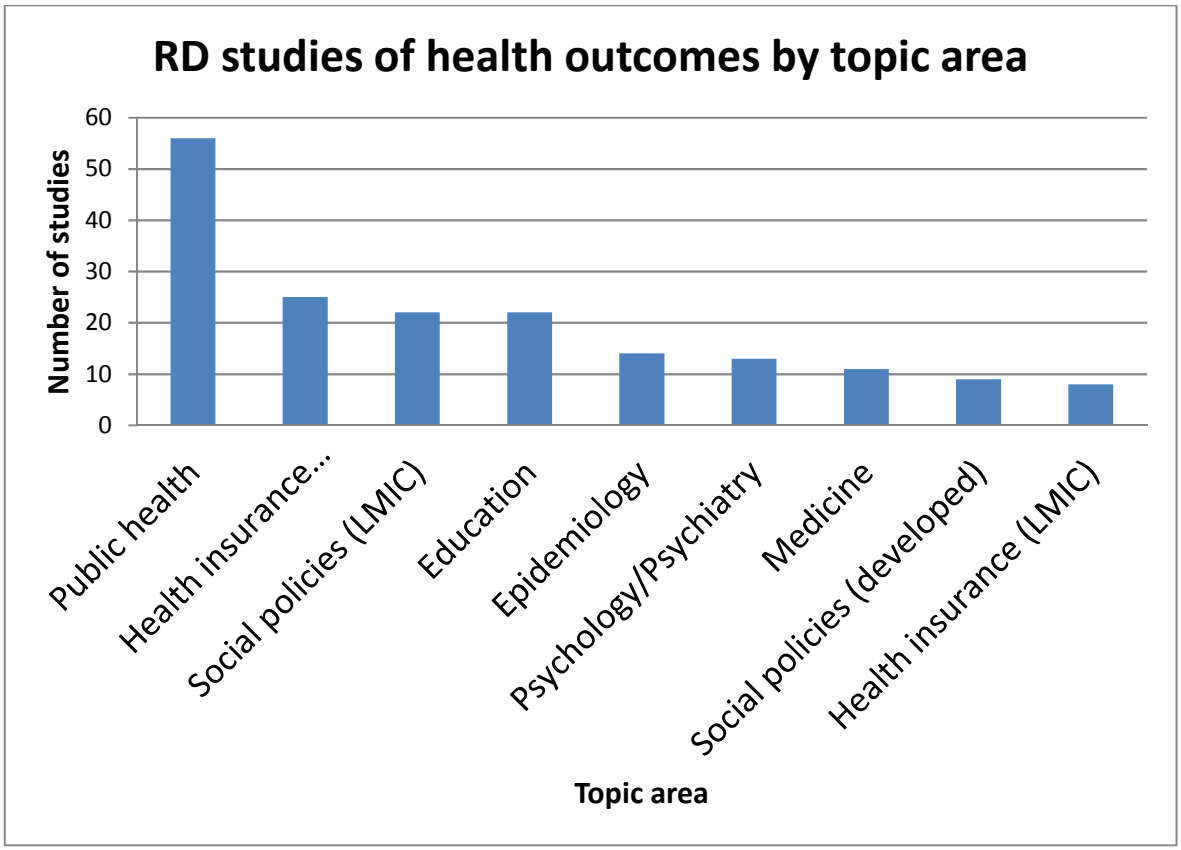
More than two thirds (124/181; 68.5%) of the RD publications identified appeared in peer-reviewed journals. Approximately one third were identified from grey literature sources. Almost half (80/181; 44.2%) of publications appeared in journals indexed in Web of Science as economics or health economics journals (Figure 3.3). Journals indexed as “Public, Environmental & Occupational Health” were the source of 14 (7.7%) of included studies.

Figure 3.3. Frequency of publications using regression discontinuity designs by academic discipline



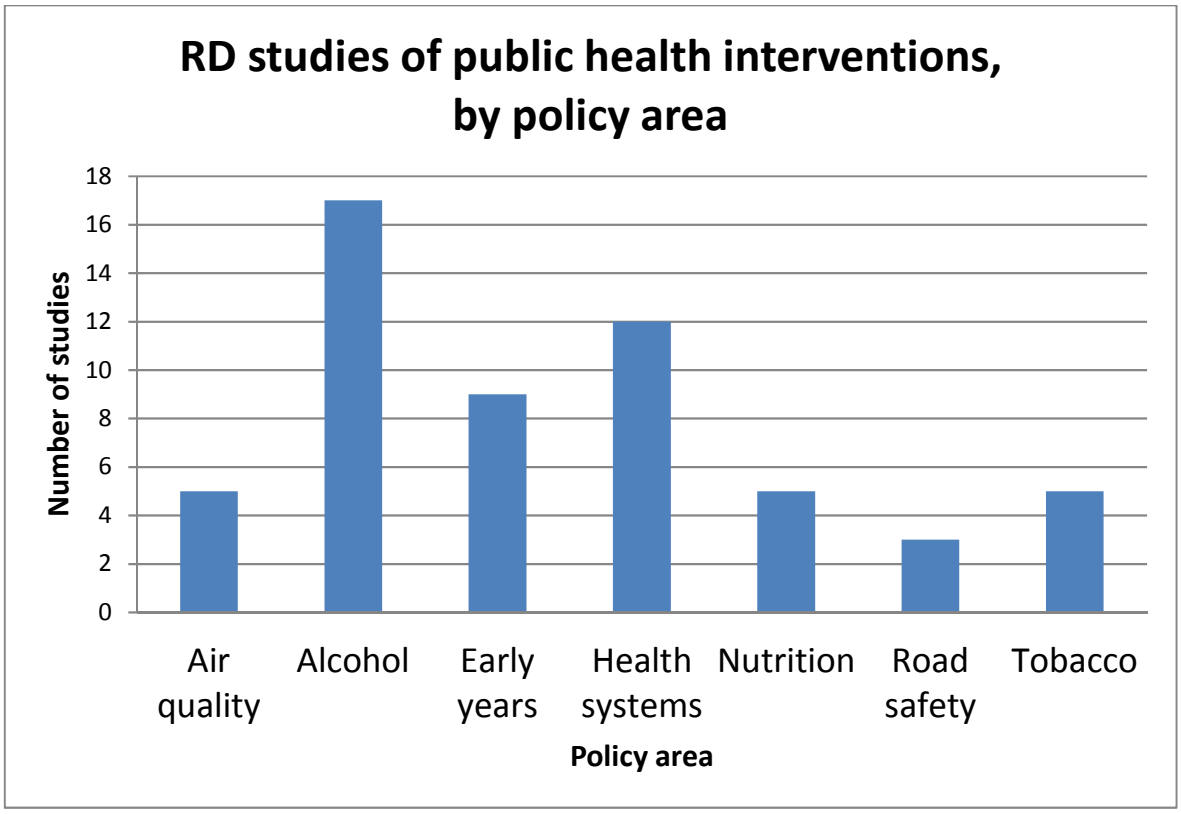
Of the included studies, nearly one third (57/181) investigated public health policy-related questions and nearly one fifth (33/181) evaluated health insurance schemes in either developed or low and middle income countries (LMIC). Figures 3.4 and 3.5 show the number of studies by topic area.

Figure 3.4. Regression discontinuity studies of health outcomes by topic area
The topic areas were identified through thematic analysis of the interventions or exposures and settings investigated.



The remaining studies considered questions of clinical effectiveness, epidemiological cause and effect, and the health effects of insurance schemes, social programmes, and education. A large number of studies (n=17) evaluated the impact of minimum legal drinking age (MLDA) legislation. The MLDA studies represent the largest number of RD studies on the same policy issue. Further analysis and synthesis of these studies form the basis of the next chapter of this thesis.

Figure 3.5. Regression discontinuity studies investigating specific public health policy topics
The topic areas were identified through thematic analysis of the interventions or exposures under study.



Tables describing the detailed characteristics of all included studies along with references to all the studies appear in Appendix 2, organised by the topic headings that appear in Figures 3.4 and 3.5 above.

A fundamental requirement of a regression discontinuity design is the presence of a forcing variable. Thematic analysis of the forcing variables used in the included studies suggested that six types of forcing variables are used in studies of health outcomes. These types can be summarised as: age; social measures such as poverty indices, literacy rates, or income; clinical measures that act as a threshold for intervention; environmental measures; geographical boundaries; and dates of events that trigger a change in exposure status, such as policy changes or disasters. Table 3.2 provides examples.

Table 3.2. Thematic analysis of forcing variables used in RD studies of health outcomes

Type of forcing variable	Number of studies	Measurement used	Threshold rule
Age	65	Age in days, months, weeks, or years	Age threshold for: • Starting school • Leaving school • Legal drinking age • Insurance eligibility • Retirement age
Date	56	Calendar date, month, or year	Dates of: • Implementation of policy/ legislation • Repeal of policy/ legislation • Disaster or major incident • Change in situation or conditions
Socioeconomic measure	39	• Company payroll total • Dropout risk score • Family income • Household acreage • Investment cost • Poverty or literacy rate • Poverty or welfare index • Predicted probability of borrowing microcredit • Programme quality score • Vote share or margin	• Benefit or programme eligibility • Election outcome • Legislated threshold
Clinical measure	10	• Addiction severity measure • Birthweight • Cardiovascular risk • CD4 count • Down syndrome risk • Exeter Alcohol Scale • Positive Symptoms Scale • PTSD Reaction Index • Systolic blood pressure • Time of birth • Weeks of gestation	• Risk threshold for intervention • Guideline threshold for intervention • Legislated threshold for intervention
Environmental measure	5	• Ozone forecasts • Air pollution levels	• Policy threshold for action • Legislated threshold for action
Geographical location	4	• Political boundary • Distance from boundary • Latitude and longitude	• Programme eligibility
Other	3	• Class size • Number of schools • Draft lottery number	• Policy threshold for intervention/exposure • Programme eligibility

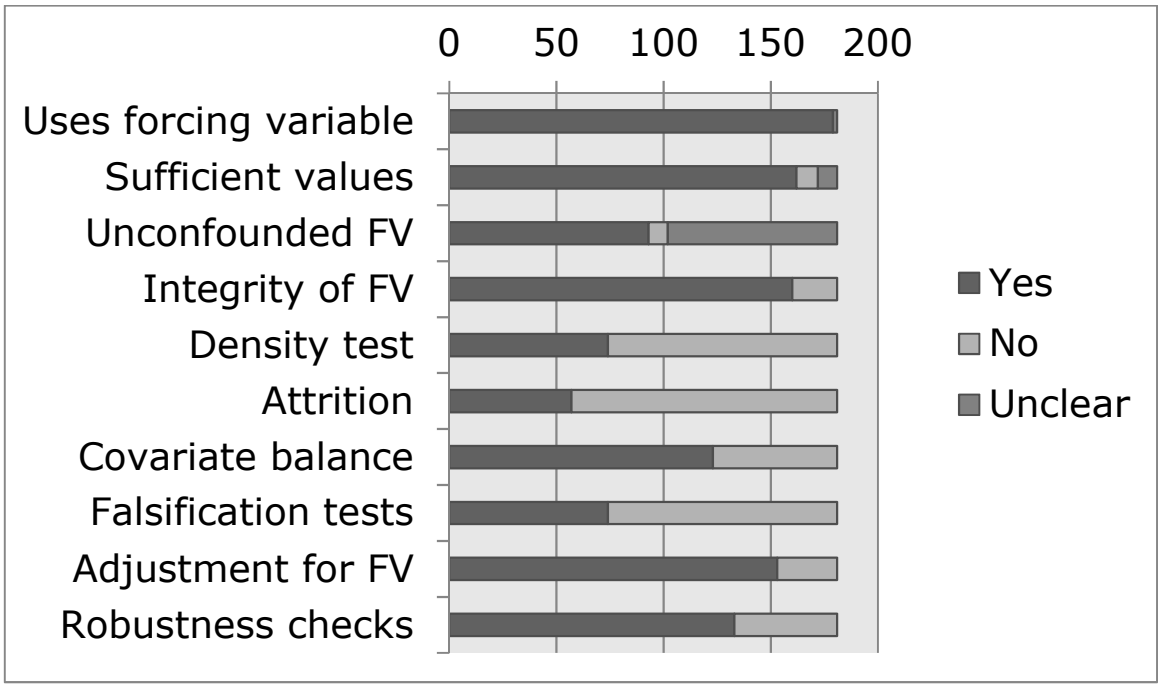
For a regression discontinuity design to be used, the forcing variable must be implemented in the context of the application of a threshold rule to assign

people or study units into treatment and control groups. Recognition of threshold rules is therefore essential to the use of RD to analyse natural experiments. Thematic analysis suggested that four sources of threshold rules are common to the included studies: programme eligibility rules for social programmes and other complex interventions; clinical decision-making rules or guidelines; thresholds imposed by legislation to restrict or limit activities that affect health; and dates of the implementation of changes to these rules. Table 3.2 provides examples.

3.5.2 Quality assessment

Study quality was assessed against ten appraisal criteria developed for this review (see chapter 5). The ten quality appraisal criteria were fully met by only 5% (9/181) of the studies. Common issues in study quality included lack of information about study attrition, failure to assess baseline equivalence on covariates, lack of density tests and falsification tests, and failure to establish that the forcing variable was unconfounded (Figure 3.6). Only 8% (15/181) of studies reported a pre-specified primary outcome or study protocol.

Figure 3.6. Summary of quality assessments of regression discontinuity studies of health outcomes
Each horizontal bar shows the number of studies (total=181) judged as yes, no, or unclear as to whether they meet ten criteria derived from the What Works Clearinghouse standards.



Almost all studies (179/181; 98.9%) clearly reported the forcing variable used (criterion 1) and most (162/181; 89.5%) reported the use of at least four discrete values of the forcing variable on either side of the cut-off value (criterion 2). In the included studies, 93/181 (51.4%) provided enough information to support a conclusion that the forcing variable was not confounded, 9/181 (5%) used a cut-off that was clearly used to assign people to additional treatments other than the one under investigation, and 79/181 (43.6%) used a forcing variable that could conceivably be confounded without reporting clear evidence to the contrary (criterion 3).

Of the included studies, 160/181 (88.4%) provided some account of scoring and treatment assignment (criterion 4), and 74/181 (40.9%) reported a density test or histogram of the forcing variable (criterion 5).

Reporting of study attrition was the area of poorest quality in these studies, with 57/181 (31.5%) reporting any information on attrition (criterion 6). Just over two thirds of studies (123/181; 68%) examined whether treatment and control groups showed baseline equivalence on any covariates (criterion 7), but less than half (74/181; 40.9%) conducted falsification tests (criterion 8).

Finally, regarding the quality of the statistical analysis, the model was adjusted for the forcing variable (criterion 9) in most, but not all, studies (153/181; 84.5%). Robustness checks of the model were reported in nearly three quarters (133/181; 73.5%) of studies (criterion 10).

3.6 Discussion

This review has identified 181 studies that apply the RD design to investigate health-related research questions, approximately six times the number of studies identified by Moscoe et al. in their 2015 review of RD despite using the same inclusion and exclusion criteria. Furthermore, the included studies cover a wide range of health and social interventions, exposures, and policy topics. Thus, this review demonstrates that RD has been applied more often and for a greater diversity of health-related research questions than was previously appreciated. The findings confirm and lend weight to the arguments of previous authors that RD is a suitable design for consideration in the evaluation of health

interventions and health policy. This review also provides some evidence against the criticism that RD and natural experiments depend upon contrived research questions that fit the available data rather than addressing genuine and meaningful evaluation problems (Dunning, 2012).

The difference in findings between the present and previous reviews indicates the importance of searching multiple databases for any systematic review, but particularly for review questions that are interdisciplinary in nature (Petticrew and Roberts, 2006). The sizeable numbers of included studies found in economics journals and grey literature suggest that systematic reviews of public health policy topics would be more comprehensive if databases such as Econlit and RePeC were included in search strategies.

This review offers two important contributions to the literature concerning policy evaluation and natural experiments. First, it offers a comprehensive view of where to look in health policy and practice for the threshold rules and forcing variables that can be exploited for analysis using regression discontinuity designs. Second, it shows the strengths and weaknesses of the existing literature in terms of study quality. Users of this review who intend to design or fund RD studies should note the variation in study quality and use the results to learn from examples of good practice and the potential pitfalls of misapplication of the design. Studies such as the evaluation of Head Start (Ludwig and Miller, 2007), for example, provide a full account of the choice of forcing variable and how it was implemented in the context of the programme; explore the sensitivity of their results to bandwidth choice; apply both parametric and non-parametric methods; and investigate and rule out rival hypotheses. Other studies demonstrate that the mere existence of a cut-off score does not necessarily make the application of RD feasible or logical. Indeed, it was apparent that some studies have misapplied the RD design in ways that violate its assumptions and potentially do not support the aims of the evaluation or the conclusions of the study.

In conducting this review, some limitations of the regression discontinuity design became evident. Previously the chief limitation of the design was recognised as the need for large sample sizes to achieve adequate statistical power. Many of the RD studies examined in this review used very large datasets and thus sample

size was less of a concern. However, in exploring functional form and conducting robustness checks in the absence of a study protocol or primary outcome, many studies inadvertently created problems for both interpretation and synthesis. RD studies frequently present the results of multiple analyses, including different stratification of data (for example, by gender or age), different choices of bandwidth around the forcing variable, and different statistical methods. Many studies have multiple outcomes without being powered for a particular primary outcome and have dozens of datapoints that could be extracted. The critical appraisal tool did not help to distinguish between studies that perform multiple analyses in a design-driven manner according to a protocol developed a priori versus apparent data dredging in which results are reported at multiple levels of significance testing, few results are statistically significant (but those that are statistically significant are cherry-picked for emphasis), and no adjustment has been made for multiple comparisons. As a result, extracting outcome data from the studies was problematic. It would also be difficult to accurately and meaningfully summarise the conclusions of such studies for decision-makers.

This review joins a small number of other systematic reviews that have investigated the application of innovative non-randomised study designs and methods to medicine, epidemiology, and public health. Compared to the findings of a systematic review of instrumental variables in epidemiology and medicine (Davies et al., 2013), more examples of RD than of instrumental variables can be identified, suggesting that, although good instruments may be hard to find, good forcing variables may be less so. These findings also support the conclusion of Moscoe et al. (2015) that RD is probably underutilised in health research: although numerous relevant applications of the design can be identified, few have been replicated or extended to other contexts, and the results suggest that the potential to do so exists. Also, RD is not yet as commonly applied as, for example, propensity score matching has been in medicine; a systematic review on that topic identified 296 studies published in a six-month period in PubMed alone (Ali et al., 2015). Finally, this review found variation and weaknesses in the quality of RD studies; the reviews of instrumental variables and propensity score matching similarly found important weaknesses and gaps in the reporting of those study types, suggesting that researchers using these methods, relatively new in medicine, epidemiology, and public health, would benefit from tools and

educational opportunities designed to promote the rigorous design, analysis, and reporting of natural experiments and other non-randomised studies.

The limitations of the review include the double-sifting of a sample rather than the full set of initial search results; however, the sample was randomly chosen and reviewer agreement was 100% after discussion. Similarly, although the critical appraisal method was piloted with two reviewers and a 10% sample of the full results was also appraised by two reviewers, with 100% agreement after discussion, the bulk of the critical appraisal results reflect assessments by a single reviewer. In both of these cases, the unexpectedly large number of included studies and limitations of time and resource prevented the involvement of two reviewers at all steps.

3.7 Chapter summary

This review contributes to the literature by identifying 181 RD studies, describing their findings, critically appraising their quality, and grouping them by policy area or clinical topic in order to facilitate either replication or the identification of opportunities for new and original research. The key strength of the review is its exhaustive search of 32 databases from multiple disciplines, including education, economics, environmental science, and sociology as well as health, as well as grey literature sources and handsearching of included papers, to provide the most systematic and comprehensive review to date of the use of regression discontinuity designs in public health, epidemiology, medicine, healthcare, and related policy areas.

The next chapter performs a further synthesis of the largest subset of RD studies on a single intervention or policy topic identified in chapter 3, namely 17 studies of minimum legal drinking age (MLDA) legislation. This further synthesis was anticipated as part of the published protocol for the RD review. Examining this subset of studies allows more detailed exploration of how a natural experimental design can be applied to answer questions of policy effectiveness, how the resulting data can be synthesised, and what challenges natural experimental studies and designs may present for systematic reviews.

4 Effectiveness of minimum legal drinking age (MLDA) laws in preventing alcohol-related harms: a systematic review

4.1 Chapter overview

Minimum legal drinking age (MLDA) laws constitute a natural experiment that is suitable for RD analysis because the drinking age threshold creates a sharp difference in alcohol availability between two groups. This chapter reports a systematic review of RD studies of MLDA, conducted within the wider review of RD studies reported in chapter 3. This chapter first places MLDA within the broader context of alcohol control policies and demonstrates the importance of these policies to public health. Then the characteristics and quality of the included studies are described. A narrative synthesis is conducted and the results visualised using an effect direction plot. Finally, the implications of the review for alcohol policy, systematic review methods, and reporting of RD studies are discussed.

4.2 Aims

By analysing and synthesising MLDA studies identified within the systematic review of RD designs in public health reported in the previous chapter, this chapter aims to investigate the following research questions:

1. How have RD designs been implemented in research on the health effects of MLDA legislation?

2. What is the evidence from RD studies on the effectiveness of MLDA legislation in reducing alcohol-related harms in young people?

3. What is the quality of RD studies on MLDA and what are the strengths and limitations of this evidence?

4. What issues do RD studies present for data extraction and synthesis in systematic reviews and how might these issues be resolved?

4.3 Background

4.3.1 Prevention of alcohol-related harms

Alcohol is a serious public health problem, causing an estimated 2.5 million deaths per year worldwide; alcohol use is a leading risk factor for premature death and disability and one of the top four modifiable risk factors for non-communicable diseases (World Health Organization, 2010). Alcohol has a causal role in breast, liver, colon, oral, and oesophageal cancers (The Lancet, 2017). The global burden of disease due to alcohol in 2004 amounted to 3.8% of all global deaths, 4.6% of the total global disease and injury burden, and 36.4% of all neuropsychiatric disability-adjusted life years (DALYs) (Rehm et al., 2009). In addition to physical and psychological harms, the social harms attributable to alcohol consumption are considerable, including domestic violence, child abuse and neglect, negative impacts on work and education, public disorder and safety issues, and crime, amounting to social costs estimated between 1 and 3 % of gross domestic product in Europe (Klingemann, 2001).

Europeans consume the most alcohol and have the highest burden of associated cancers (The Lancet, 2017). In the UK, some ten million adults exceed the recommended maximum intake of 14 units per week (Williams et al., 2018); in Scotland, the equivalent of 19.6 units of alcohol per adult were sold per week in 2017 (NHS Health Scotland, 2018). Alcohol-related hospital admissions exceeded 24,000 in 2016-17 and demonstrated pronounced inequalities, with rates of stay more than eight times higher in the most deprived compared to the least deprived areas in Scotland (NHS Health Scotland, 2018).

The multifaceted disease and societal burden associated with alcohol combined with the scope and magnitude of harmful alcohol consumption suggests that policy action is imperative; however, both the WHO (Casswell and Thamarangsi, 2009) and the United Kingdom (Williams et al., 2018) have been criticised for inadequate policy responses. Like smoking, alcohol consumption is a complex social behaviour which involves vested economic interests and which can potentially be addressed through a variety of programmes, policies, and legislation at the individual, health service, and population levels. The World Health Organization Global Strategy to Reduce the Harmful Use of Alcohol (2010)

grouped policy responses into ten recommended target areas, which included leadership, community, and health service responses; policies to target drink-driving, alcohol availability, pricing, and marketing; harm reduction; addressing illicit alcohol production; and monitoring and surveillance.

In Scotland, an alliance of organisations led by Alcohol Focus Scotland produced a series of recommendations to inform the next iteration of the Scottish Government's alcohol strategy (Alcohol Advocacy Coalition, 2017). The recommendations included: a Health in All Policies approach; pricing and taxation reforms; restricting availability through licensing and enforcement; changes to marketing and labelling; health promotion actions; and improvements to healthcare and social services. Given the importance of the problem to public health and the variety of policy options available, there is arguably an ongoing need for evidence synthesis to inform policy decisions.

An abundance of systematic review evidence exists to support decisions in alcohol policy. Previous systematic reviews have addressed the effectiveness of different alcohol policy approaches and interventions, including pricing, taxation, licensing, labelling, and marketing restrictions. A systematic overview and synthesis of these reviews is beyond the scope of this chapter. The following section will summarise a selection of systematic reviews that set the scene for the present work by providing a global overview of policy effectiveness, specifically considering the contribution of natural experiments, and more closely examining policy effectiveness in the UK context. The focus then shifts to reviews that address minimum legal drinking age legislation.

Two overviews of systematic reviews have demonstrated that most alcohol policy interventions are supported by evidence of effectiveness and meet thresholds of cost-effectiveness. Anderson, Chisholm, and Fuhr (2009) conducted an overview of systematic reviews and meta-analyses of policies to reduce alcohol-related harms. They noted that the conceptual framework and theoretical basis of such policies (such as deterrence and cost increases) is well understood and generally applicable across societies. Their narrative synthesis was structured according to the target areas of the WHO Global Strategy and identified evidence to support policy effectiveness in all areas apart from education, community programmes, harm reduction in bars, and illicit alcohol

production. They also considered cost-effectiveness and determined that, in Europe, population policy approaches (drink-driving legislation and enforcement, reduced retail access, advertising ban, and pricing policies) were more cost-effective than health sector interventions, ranging from I$ 335 to 961 (international dollars) per DALY saved (the cost per DALY for brief clinical interventions for heavy drinkers, by comparison, was I$2671) (Anderson, Chisholm, and Fuhr, 2009).

In 2013, a similar but more methodologically rigorous overview of systematic reviews was reported by Martineau et al., who produced a narrative synthesis of 52 reviews, 12 of which were high-quality (Martineau et al., 2013). Their findings on effectiveness mirror those of Anderson et al. (2009): consistent evidence that taxation, drink-driving policies, policies to restrict sales availability, and mass media campaigns were effective, mixed or weak evidence to support interventions in family, educational or workplace settings, and a lack of evidence on harm reduction in bars, illicit alcohol interventions, and community interventions.

The broadly positive picture of evidence for the effectiveness of alcohol policy interventions may be somewhat different if study selection criteria are changed. Nelson and McNall conducted a systematic review of pricing and tax policies evaluated as natural experiments (Nelson and McNall, 2016, Nelson and McNall, 2017). They argued that natural experiments should be of particular use in evaluating the causal effects of policies, but noted these study types have been neglected in previous systematic reviews. They identified 45 studies that assessed the effects of policy changes on alcohol-related harms in nine countries (2016). Contrary to Anderson et al., they found a mixture of positive, null, and negative effects. They found a similar mixture of effect directions when considering alcohol consumption and drinking patterns as outcomes (2017). For these outcomes, 29 papers from five countries were identified and almost all used survey data to construct regression models to evaluate policy effects. Nelson and McNall concluded that the evidence base was inconsistent and insufficiently robust to inform policy development.

An additional review has systematically examined and synthesised evidence on these same alcohol control policies, but with specific reference to implications

for policy and public health professionals in the UK. Commissioned by the Department of Health, Burton et al. conducted a rapid review of studies published between 2000 and 2016, organising alcohol control policies into seven areas broadly similar to other reviews, although they also included brief interventions and treatment in healthcare, workplace, and criminal justice settings as an alcohol control policy area (Burton et al., 2017). This review was innovative in applying GRADE methodology, modified so that the hierarchy of evidence included natural experiments and modelling studies, to assign a strength of evidence rating to each policy intervention. This review also differed from others discussed above by including minimum unit pricing (MUP), citing UK modelling studies and natural experiments from Canada as evidence. The review concluded that policies that reduce the affordability of alcohol are the most effective and cost-effective, and that there is strong evidence to support regulation of marketing. Although the evidence to support drink-driving legislation was graded 'high' and found to be both effective and cost-effective, the review concluded that in England such legislative measures "are estimated to lead to minimal public health gains compared with policies such as taxation. Nonetheless, reducing drink-driving is an intrinsically desirable societal goal" (p. 1574). The review did not include minimum legal drinking age among the alcohol control policies investigated.

Mapped against the range of policy interventions investigated in these systematic reviews, current areas of alcohol policy development in the UK may seem relatively limited. In Scotland, the introduction of a minimum unit price of 50p per unit of alcohol has been an important policy action to address alcohol-related harm through a population-level intervention, the expected effectiveness of which has been supported by the findings of a systematic review (Boniface et al., 2017). In England, the 2010-2015 Coalition Government consulted on a new alcohol strategy in 2012 and in 2013 published its response to the consultation (Home Office, 2013). Then-Home Secretary Theresa May argued that MUP should be delayed until "conclusive evidence" of effectiveness was available and to prioritise engagement with the alcohol industry instead of using "the sledgehammer of national legislation, which often misses its target" (Home Office, 2013, p.7). The effectiveness of interventions incorporated into the resulting Public Health Responsibility Deal was subsequently examined in an

overview of systematic reviews (Knai et al., 2015), which concluded that the Responsibility Deal was largely based on information and communication interventions that are probably ineffective at changing behaviour.

Although comprehensive overviews of evidence are not lacking, comprehensive policy is more difficult to find. The Responsibility Deal ceased in 2015 with the change of Government but has not to date been replaced with a coherent framework or strategy for alcohol policy in England. Similarly in Scotland, the last comprehensive alcohol strategy, the Alcohol Framework for Action, was published in 2009 and a "refresh" of the framework, promised for early 2018, has at the time of writing (September 2018) not yet materialised (Scottish Government, 2018).

4.3.2 Minimum legal drinking age legislation

Policies to prevent alcohol-related harm may operate universally, by reducing risks at a population level (for example, advertising bans), and/or selectively, by targeting groups who are disproportionately at risk of alcohol-related harms (for example, minimum unit pricing, which aims to reduce consumption in heavy drinkers). Young people face a particular risk of harm due to alcohol's neurotoxicity, which can adversely affect brain development (Scottish Health Action on Alcohol Problems, 2014, Spoth et al., 2008). Adolescent drivers are also at heightened risk of motor vehicle accidents (MVA), which are the leading cause of death for people aged 16-19 in the United States, and drinking any amount of alcohol increases the risk of MVA in this age group compared to older drivers (Centers for Disease Control and Prevention, 2017). Policy interventions to prevent harm in this group include driver licence restrictions (such as graduated licence programmes) and age restrictions on alcohol sales and consumption. The most well-studied of such policy interventions is the minimum legal drinking age (MLDA) (Wechsler and Nelson, 2010).

MLDA laws have been in place in the United States since at least 1933 and, after states that lowered their MLDA to 18 years old were found to have higher MVA rates, an MLDA of 21 years old was in place across the country by 1988 (Wechsler and Nelson, 2010). At least two comprehensive systematic reviews of MLDA in the United States have concluded that these laws are effective in reducing

alcohol-related harms. In 2002 Wagenaar and Toomey comprehensively reviewed 132 studies that evaluated the MLDA from 1960 to 1999 (Wagenaar and Toomey, 2002), a work that has been described as “definitive” (DeJong and Blanchette, 2014). They evaluated the quality of these studies according to three criteria of sampling design (probability sampling or census data = higher quality), study design (pre-post, longitudinal, and time series higher quality compared to cross-sectional), and presence of a comparison group. They coded study results according to their direction of effect and statistical significance. They argued that the preponderance of evidence, particularly of higher-quality studies with statistically significant effect estimates, showed that a higher MLDA reduced alcohol consumption and MVA; that evidence on other outcomes such as suicide and vandalism was inconsistent; and that lack of enforcement was a mediating factor in effectiveness (Wagenaar and Toomey, 2002). Their review was well reported, with transparent and reproducible methods, only lacking detail in how final judgments were made from a complex synthesis, which appeared to rely on vote-counting of statistically significant effects.

At the same time the Task Force of Community Preventive Services was undertaking a related systematic review on behalf of the CDC and National Highway Traffic Safety Administration (Shults et al., 2001). They evaluated the effectiveness of five policy interventions to reduce alcohol-related MVA. For MLDA they included 33 studies that investigated the effects of changing the MLDA in the USA, Canada, and Australia. They included only time series or controlled before and after designs. They concluded there is “strong evidence” that MLDA laws are effective in preventing alcohol-related MVA and related injuries (p. 75).

Despite this evidence base and apparent consensus, the MLDA became a subject of renewed debate in America with the launch of the Amethyst Initiative, a campaign to lower the MLDA organised by some university and college presidents and chancellors as their observations of underage drinking on campuses led them to believe existing legislation was not effective in regulating behaviour (Amethyst Initiative) (n.d.). This campaign came as a “surprise” in public health and road safety circles “given an extensive research literature showing that the age 21 MLDA reduces injuries and saves lives” (DeJong and Blanchette, 2014) and

led to renewed interest in research on MLDA and the underlying evidence base. In a commentary in the American Journal of Public Health, Wechsler and Nelson argued that public health professionals needed familiarity with the evidence base on MLDA in order to "advocate effective public policy" (2010, p. 988). They note that debate has centred on two questions: whether the MLDA actually had a causal effect, and whether lowering the MLDA from 21 to 18 would actually change the behaviour of 18-20 year olds (2010, p. 989).

Both questions can be addressed by identifying and analysing the many natural experiments that have arisen as MLDA legislation has been introduced and revised in different jurisdictions over time. RD is well suited for such analyses because two types of thresholds can arise that sharply divide people into exposed and unexposed groups without any opportunity for them to interfere with their allocation. One such threshold is the date on which legislation is enacted. The other is the drinking age limit itself. People who age past that threshold gain legal access to alcohol and thus experience what is effectively a price decrease in the total cost of obtaining alcohol to the individual. People just above the age limit can be compared to those just below. Given the assumption that all other characteristics that could affect alcohol-related outcomes are smoothly distributed across the threshold, any differences in outcomes between the groups can be causally attributed to legal access to alcohol. Carpenter and Dobkin (2009, 2011) were the first to identify MLDA as a natural experiment that could be analysed using RD. In an introduction to econometric methods, Angrist and Pischke cite Carpenter and Dobkin's MLDA work as a paradigmatic example of the design, remarking that their studies "appear to have been written in RD heaven" (Angrist and Pischke, 2015, p. 164). The insensitivity of the results to specification or bandwidth, they argue, "suggests the findings generated by an RD analysis of the MLDA capture real causal effects" (p. 164).

4.4 Methods

4.4.1 Identification and appraisal of studies

The RD MLDA studies were identified within the larger systematic review of RD studies in health described in chapter 3, where the methods are reported in full.

In brief, the search encompassed 32 databases, using “regression discontinuity” or “regression-discontinuity” as search terms in title, abstract, or full text (when available), as well as the reference lists of included papers. Studies published between 1960 and March 2015 in any language were included that used a regression discontinuity design to investigate any physical or mental health outcome for any intervention or exposure. No search filters were used as none have been developed for RD studies.

The citations retrieved were downloaded into an EndNote database. Titles and abstracts were screened by one reviewer (myself); a 10% random sample was screened by a second reviewer (HT) and the results compared. Full text papers were then screened for relevance by one reviewer (myself); a 10% random sample was screened by a second reviewer (HT) and the results compared. Disagreements were resolved through discussion and involvement of a third reviewer when required.

The RD studies thus retrieved were then categorised by topic area. The review protocol stated that a topic area in which several studies were identified would be the subject of more detailed analysis. Seventeen studies used an RD design to evaluate the health effects of MLDA legislation, the highest number of studies on a single topic. These studies were therefore chosen as the subject of further investigation for critical appraisal, narrative synthesis, and possible meta-analysis.

At the time of this review, no critical appraisal tools for the quality assessment of RD studies had been published in the health research literature or by evidence synthesis organisations such as the Cochrane Collaboration. The only quality assessment tool identified through the literature search was a set of standards for the evaluation of RD studies to be used as evidence in education policy and planning decisions. This tool, the What Works Clearinghouse Standards for RD, is described in detail in chapter 5. Two reviewers (myself and MC) independently assessed each included study, recorded the assessments, and met to discuss and agree a final assessment for each study.

4.4.2 Data extraction

Data were extracted to describe each study in terms of the country and time period represented, the natural experiment under investigation, the forcing variable used, number and summary characteristics of participants, outcomes examined, primary outcome if stated, data sources used.

In order to consider the validity of the RD approach, data were extracted on any other approaches used to analyse the same natural experiment within the study (for example, panel data, time series, or IV) and any efforts at falsification of the RD approach.

In order to investigate the nature of the information provided within RD studies and the potential challenges presented for systematic reviews, data were also extracted that described the characteristics of the statistical analyses presented in these studies. The extracted data included: whether the analysis was based on a pre-established protocol, the number of models reported, variables included in models, subgroup analyses performed, the model selection method, whether this selection was made a priori, a description of the preferred model, and the number of observations in the preferred model compared to the full sample.

Finally, outcome data were extracted (where available) for mortality (all cause, motor vehicle related, and suicide), alcohol-related hospital admissions, and motor vehicle accidents. These outcomes were selected because of their relevance to public health policy given their direct and high costs to both individuals and society, and because they can be interpreted (within the context of the RD design) as a direct measure of the effects of MLDA on health.

4.4.3 Synthesis methods

The design and methods used in the included studies to analyse MLDA are presented as a narrative synthesis (Popay et al., 2006). Aspects of study context and details of statistical approaches are presented in tables.

If event rate data were available from two or more studies for a given outcome, the protocol specified that a meta-analysis would be performed. Meta-analysis

was considered for two outcomes, change in rates of all-cause mortality and change in rates of mortality due to motor vehicle accidents, which were reported in a comparable statistical manner in two studies (Carpenter and Dobkin, 2011; Carpenter, Dobkin, and Warman, 2014). These studies reported data for these outcomes as a rate difference with standard error, but without the numbers of events observed on either side of the threshold nor the denominators used, meaning that any recalculation of the outcome was not possible. Accordingly meta-analysis could not be undertaken.

As the meta-analysis was not possible, and at best could only have incorporated two of the 17 studies, a third synthesis approach was applied to make the best use of the available evidence. Estimates of the effect of MLDA on mortality, hospital admissions, and MVA were synthesised in the form of effect direction plots (Thomson, 2013, Thomson and Thomas, 2013). Effect direction plots provide a visual summary of a body of evidence for a given systematic review, showing the included studies and relevant outcomes in a grid along with symbols for the direction (increased or decreased effect or risk) and statistical significance of each effect estimate from each study. This visual summary complements a detailed and complex narrative review, from which it may be difficult to get an overall sense of the evidence or to draw conclusions.

Some modifications have been made to the effect direction plot. The example plot shown in the original methods papers (Thomson and Thomas, 2013; Thomson, 2013) is taken from a Cochrane review on housing improvement interventions and included columns for study design, time since intervention, and intervention integrity. These columns have been omitted as they are not necessary to describe the results of the present review. Study quality has also been omitted as current methods do not support summarising the quality appraisal of RD studies in a single letter or symbol. Numbers in sample has been omitted as it is an area of incomplete reporting in these studies and the relative sample size of the studies can be represented graphically.

The most important modification relates to synthesis of multiple outcomes. The effect direction plot was originally conceived to support the merging of heterogeneous yet conceptually related outcomes into a single outcome category. For example, diverse measures such as cough frequency, cough

severity, wheeze incidence, wheeze duration, and asthma exacerbations could be represented under one category, “Respiratory”, and effects synthesised despite heterogeneity. This approach allows the end user to form a judgment as to whether the intervention improves conceptually related outcomes and is useful when the judgment does not depend upon a point estimate or effect size for a precisely defined clinical endpoint. The approach provides a solution when a body of studies evaluate a similar intervention but use diverse outcome measures.

In the MLDA RD studies, the chief difficulty in synthesis is not so much a diversity of outcome measures as a diversity of modelling specifications, which results in multiple effect estimates for each outcome measure (as reported in section 4.5.4 below). Accordingly, for the effect direction plot to be useful for RD studies, it must allow not only for synthesis of multiple outcomes but also for synthesis of multiple specifications. To this end, the original methodology has been adapted to provide synthesis rules to account for situations in which direction of effect and statistical significance vary across model specifications. The original methodology specified decision rules based on the percentage of outcomes in the study reporting a consistent direction or statistical significance of effect (60% or 70%, depending on the rule). The number of model specifications to be synthesised per outcome ranged from 1-9 in the present review with a median of 6. A pragmatic decision was made to change the decision threshold to 2/3 of specifications as it was better suited to the data at hand. Furthermore, studies reporting only one specification and only one outcome were to be flagged with an asterisk (*) to indicate that they were not subject to the synthesis decision rules. Otherwise, the effect direction plot could be incorrectly interpreted such that consistency of effects in these studies could be overestimated.

4.5 Results

4.5.1 Included studies

Of the 181 RD studies identified in the systematic review, 17 investigated the health effects of minimum legal drinking age legislation. MLDA was the most frequently assessed intervention or exposure among RD studies of health

outcomes. Table 4.1 reports the study characteristics in terms of the setting, dates, and natural experiment analysed. One study examined the MLDA in Australia (Lindo et al., 2014), two in New Zealand (Boes and Stillman, 2013, Conover and Scrimgeour, 2013), five in Canada (Callaghan et al., 2014b, Callaghan et al., 2013a, Callaghan et al., 2013b, Callaghan et al., 2014a, Carpenter et al., 2014), and nine in America (Carpenter and Dobkin, 2009, Carpenter and Dobkin, 2011, Carpenter and Dobkin, 2015b, Crost and Guerrero, 2012, Crost and Rees, 2013, Deza, 2013, Ertan Yoruk and Yoruk, 2015, Ertan Yörük and Yörük, 2012, Yörük and Yörük, 2011).

Table 4.1. Characteristics of regression discontinuity studies of the health effects of minimum legal drinking age legislation

Study	Country	Dates Represented in Data	Natural Experiment	Forcing Variable
Boes and Stillman 2013 (date-based)	New Zealand	1996-2007	Policy change (SLAA1999) which lowered the MLDA to 18	Date of policy change (December 1999); monthly data
Boes and Stillman 2013 (age-based)	New Zealand	1996-2007	MLDA of 20 (pre-law change) or 18 (post-law change)	Age (in quarters)
Callaghan et al. 2013a	Canada	April 1997-March 2007	MLDA of 18 or 19 (province/territory dependent)	Age in months (range 72 except MVA which is 48)
Callaghan et al. 2013b	Canada	April 2002-March 2007	MLDA of 19	Age in months, range 16-22 (72 months)
Callaghan et al. 2014a	Canada	1980-2009	MLDA of 18 or 19 (province/territory dependent)	Age in months, range 48 months
Callaghan et al. 2014b	Canada	2000-2012	MLDA of 18 (Québec)	Age in weeks (range 52)
Carpenter and Dobkin 2009	USA	1997-2005 (NHIS, alcohol consumption); 1997-2004 (mortality)	MLDA of 21	Age in 30-day blocks (range 19-23 years old)
Carpenter and Dobkin 2011	USA	1997-2003	MLDA of 21	Age in months, range 19-23 (48 months)
Carpenter, Dobkin, and Warman (2014)	Canada	1980-2008	MLDA of 18 or 19 (province/territory dependent)	Age in days (MLDA ±24 mos)
Carpenter and Dobkin 2015	USA	Various depending on state and outcome; 1990-2010	MLDA of 21	Age in months, range 19-23 (48 months)
Conover and Scrimgeour 2013 (date-based)	New Zealand	1993-2006	Policy change (SLAA1999) which	Date of policy change

Study	Country	Dates Represented in Data	Natural Experiment	Forcing Variable
			lowered the MLDA to 18	(December 1999); monthly data
Conover and Scrimgeour 2013 (age-based)	New Zealand	1993-2006	MLDA of 20 (pre-law change) or 18 (post-law change)	Age in days from MLDA (±11 months)
Crost and Guerrero 2012	USA	2002-2007	MLDA of 21	"Each observation is the average of substance use over a month-of-age cell" (Tables 2 and 3)
Crost and Rees 2013	USA	2000-2006	MLDA of 21	Age in months, range 19-22 (48 months)
Deza 2015	USA	1997-2009	MLDA of 21	Age in months, range 19-23 years
Ertan Yoruk and Yoruk 2012	USA	2000-2006	MLDA of 21	Age in days (MLDA ±732 days; range ages 19-22)
Ertan Yoruk and Yoruk 2015	USA	2000-2006	MLDA of 21	Age in days (MLDA ±732 days; range ages 19-22)
Lindo, Siminski and Yerokhin 2014	Australia	2000 or 2001 to 2010 or 2011	MLDA of 18 (NSW)	Age in days (MLDA ±22 mos)
Yoruk and Ertan Yoruk 2011	USA	2000-2006	MLDA of 21	Age in days (MLDA ±732 days)

Note: Boes and Stillman (2013) and Conover and Scrimgeour (2013) each report two different RD analyses, corresponding to the two different MLDA natural experiments that can be identified in New Zealand.

All of the identified studies were retrospective RD designs which used data obtained from national administrative databases or nationally representative longitudinal surveys. Each study included multiple years of data from these sources, with a mean of 11.2 calendar years (range 6 to 30) represented in the study data. In terms of the currency of the data used, the decade 1997-2007 is covered in all studies (range of dates from 1980 to 2012).

The approach to investigating the effects of MLDA was broadly similar across studies. All studies used age (in days, weeks, months, or quarters) as a forcing variable to examine an outcome in people aged just above or below the treatment threshold, i.e. the age at which they can legally purchase and consume alcohol. In the sole Australian study, this age threshold (in New South

Wales) was 18. The American studies used the age threshold of 21 which has been in place in all states since the 1980s. In Canada, the MLDA varies by province (18 in Alberta and Québec, 19 elsewhere). In New Zealand, the MLDA was lowered from 20 to 18 in 1999, creating two natural experiments which were both exploited by the two included studies. Boes and Stillman (2013) and Conover and Scrimgeour (2013) analysed both age-based and date-based discontinuities.

For the age-based RD designs, the natural experiment under investigation was the removal of legal restrictions on alcohol purchase and consumption that occurs when individuals cross the age-based threshold. In economic terms, this situation represents a discontinuous 'price decrease' for alcohol in terms of the full personal and social costs an individual may incur for consuming alcohol (Carpenter, Dobkin and Warman, 2014, p.11). This situation allowed investigation of the effects of legal access to alcohol versus age-restricted prohibition on alcohol consumption and its proximal sequelae, including motor vehicle accidents, attendance at A&E, hospital admissions, and mortality risk. Some authors additionally identified within this natural experiment the opportunity to investigate whether alcohol is a complement or substitute for marijuana and other drugs by comparing age-based discontinuities in the consumption of these substances (Crost and Guerrero 2012; Crost and Rees 2013; Deza 2015; Yoruk and Ertan Yoruk 2011).

Additionally, Boes (2013) and Conover and Scrimgeour (2013) investigated the natural experiment represented by the New Zealand Sale of Liquor Amendment Act (SLAA) 1999. The enactment of this legislation created a date-based discontinuity which the studies analysed using monthly data, with date as a forcing variable and December 1999 (when the legislation was passed) as the cut-off. This design assumes that no other changes took place at the same cut-off that would affect the outcomes. Both sets of authors acknowledge and address this issue. Boes and Stillman state their belief the assumption holds because they know of no other policy changes that occurred at that time; Conover and Scrimgeour test the assumption by implementing a difference-in-discontinuities estimator. Both acknowledge, however, that SLAA 1999 was a legislative package that involved not only lowering the MLDA but also changes to

where and how alcohol could be sold, accompanied by changes in enforcement of these laws. Therefore, any discontinuities must be seen as effects of SLAA 1999 as a whole and not exclusively of the MLDA component of the package.

The age-based RD studies would similarly be at risk of confounding and invalidation if other factors which contribute to the outcome also change at the same age cut-off. The studies address these concerns through design, analysis, narrative argument, or a combination of these approaches. Some studies examined whether a discontinuity occurred at the cut-off in an outcome that could not plausibly be caused by MLDA; such a discontinuity would serve as evidence of another factor that could be causing discontinuity in the outcomes of interest, whereas absence of a discontinuity would support the validity of the RD design. For example, Callaghan et al. (2013b), investigating whether hospital admissions in Ontario were discontinuous at the MLDA, demonstrated that rates of admission for appendicitis (which should not be affected by increased alcohol consumption) were not discontinuous at the cut-off. Less commonly, some studies provided an argument as to the plausibility of other changes at the same cut-off affecting the outcomes. In a different approach to testing RD assumptions, Lindo et al. (2014) acknowledged that the age-18 cut-off in Australia corresponds to the 'age of majority' at which young people are considered to become adults and this could confound the RD design. They addressed this issue by testing for and ruling out discontinuous changes in demographic characteristics that could serve as 'coming of age' markers, such as living at home or being employed.

4.5.2 Quality of studies

Detailed critical appraisal results for each study are presented in Appendix 3. A summary of the quality of the included studies is presented in chapter 5, figure 5.1. All studies met the three qualifying questions of the WWC Standards for RD, with the New Zealand studies deemed to meet question 3 (unconfounded forcing variable) on the basis that the studies acknowledge they are assessing the effects of SLAA 1999 as a package. All studies met the standard for integrity of the forcing variable as neither data subjects nor the researchers would have had the opportunity to alter the birth records in the datasets used. Probably because the data sources and retrospective nature of the studies precluded such

manipulation of the forcing variable, most studies did not report conducting tests of the smoothness of the forcing variable at the threshold, which led both reviewers to agree to assign a judgment of 'not applicable' to these studies for criterion 1B. However, four studies did present graphs of number of observations by age in order to present visual evidence of the smoothness of the forcing variable across the threshold, which is sufficient to meet the criterion.

Only one of the 17 studies (Deza 2015) provided any information about attrition, an area of very poor reporting in RD. Although data on attrition may not be available or may not be considered an important source of risk of bias when using comprehensive government datasets, as many of these studies do, over half (9/17) used survey data for which information on attrition rates is relevant and available. The WWC Standards require RD studies to meet the same standard for reporting attrition as randomised trials. Moreover, failure to meet the attrition standard leads to failure to meet the overall quality standard. The second reviewer and I agreed that this was not helpful for describing the quality of the studies and decided to add a judgment of 'not applicable' for retrospective RD studies.

A majority (10/17; 58.8%) of studies failed to meet the third WWC quality standard, which requires studies to verify that there are no discontinuities at the cut-off in covariates other than the forcing variable that are correlated with the outcome (criterion 3A) or in the outcome at values of the forcing variable other than the cut-off (criterion 3B). Failure to meet this standard meant that the overall judgment for that study would be 'met with reservations'.

The studies generally performed well against standard 4, which assesses the quality of the statistical modelling and reporting. Fifteen (88.2%) of the studies met all applicable criteria, while one study (Lindo et al. 2014) met with reservations because it (by not presenting graphs with fitted curves) did not meet the full criterion for graphical analysis and another (Callaghan et al. 2013a) met with reservations because it did not report results separately by province when it could have. These were minor quality issues compared with criteria 4A and 4C, which address model specification and robustness; failure against these criteria would reflect serious risk of bias in the study results, but all studies met these two key criteria.

Overall six of 17 studies (35.3%) fully met the WWC standards (modified to allow omission of density tests, on the assumption that the forcing variable could not be manipulated, and omission of reporting of attrition, on the basis that it was not reasonable to expect retrospective, population-level studies to investigate attrition to the same standard as an RCT). The remaining 11 studies (64.7%) met the WWC standards with reservations: ten because they did not conduct falsification tests of the cut-off, and one because of failure to report results separately by province. In summary, approximately one-third of the studies are at low risk of bias and two-thirds might be said to be at a moderate risk of bias because of insufficient assessment of the smoothness condition.

4.5.3 Reporting of RD analyses

Table 4.2 provides information on study reporting in relation to participants/population and outcomes. Four of 17 studies (24%) reported the exact number of participants or records analysed in the study (Carpenter and Dobkin, 2015; Crost and Benjamin, 2013; Deza, 2015; Lindo, 2014). Four studies (Carpenter 2014, Crost 2012, Ertan 2012, and Yoruk 2011) reported an 'approximate' sample size in thousands and four studies provided numbers of observed events which varied across outcomes and subgroups (Boes and Stillman, 2013; Callaghan et al., 2013b; Conover and Scrimgeour, 2013; Ertan 2015). Only one study (Boes and Stillman, 2013) specifically named one outcome measure as the primary outcome of the study.

Table 4.2 Reported number of participants, outcomes, and data sources used in RD studies of MLDA

Study	Number of Participants	Outcomes	Data Sources
Boes and Stillman 2013 (date-based)	Not reported	Alcohol-related hospital admission rates per 10,000 population; alcohol-related MVA	NZ Ministry of Health hospital episode database; NZ Ministry of Transport data on MVA; population estimates from Statistics New Zealand (denominators)
Boes and Stillman 2013 (age-based)	Not reported	Alcohol-related MVA	NZ Ministry of Transport data
Callaghan et al. (2013a)	Not reported	Alcohol-related hospital admission rates per 1,000 population	CIHI Hospital Morbidity Database

Study	Number of Participants	Outcomes	Data Sources
Callaghan et al. 2013b	Not reported	Morbidity (alcohol-related inpatient and emergency admissions)	Rates per 1000 hospital events: CIHI Hospital Morbidity Database and National Ambulatory Care Reporting System
Callaghan et al. 2014a	Not reported	Mortality: All-cause, external causes, internal causes (+/- MVA), MVA. Mortality counts within each age-in-month, NOT rates	Statistics Canada VICES (Vital Integration Capture and Edit System) capturing all deaths in Canada
Callaghan et al. 2014b	Not reported	Alcohol-related MVA	Provincial government MVA database (SAAQ)
Carpenter and Dobkin 2009	Not reported	Alcohol consumption, mortality (internal and external causes: alcohol, homicide, suicide, MVA, drugs, external other): rates per 100,000	NHIS (consumption); National Vital Statistics or NHCS (mortality)
Carpenter and Dobkin 2011	Not reported	Mortality (age-specific mortality rate, estimated) alcohol consumption	Mortality due to all causes, MVA, alcohol overdose, or suicide: estimated using National Vital Statistics records and US Census population estimates
Carpenter, Dobkin, and Warman (2014)	Approx. 36,000	Alcohol consumption, mortality (internal and external causes: alcohol, internal, external, MVA, injuries): rates per 100,000	Statistics Canada (mortality), National Population Health Surveys, Canadian Community Health Surveys
Carpenter and Dobkin 2015	N records = 3770267	Alcohol-related ED visits and hospital admissions per 10,000 person-years	Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases and State Emergency Department Databases
Conover and Scrimgeour 2013 (date-based)	Unclear. Table 1 suggests 872085 hospitalizations observed of which 2.3% (20057) were alcohol-related, but tables 5 and 6 give very different n of observations	Alcohol-related hospitalizations	New Zealand Health Information Service
Conover and Scrimgeour 2013 (age-based)	Unclear, see above	Alcohol-related hospitalizations	New Zealand Health Information Service
Crost and Guerrero 2012	Unclear; survey sample is "approximately 70,000 people"	Alcohol and marijuana use	National Survey of Drug Use and Health
Crost and Rees 2013	28,089	Marijuana use	NLSY97 longitudinal survey
Deza 2015	8984	Alcohol and hard drugs consumption	NLSY97 longitudinal survey
Ertan Yoruk and Yoruk 2012	Approx. 9000	20-point psychological wellbeing index based on Mental Health Inventory (self-	NLSY97 longitudinal survey

Study	Number of Participants	Outcomes	Data Sources
		reported); alcohol consumption	
Ertan Yoruk and Yoruk 2015	Not reported; no obs per outcome ranges from 6,999 to 26,417	Alcohol consumption and 8 sexual behaviour outcomes	NLSY97 longitudinal survey
Lindo, Siminski and Yerokhin 2014	n = 2359 (HILDA, alcohol consumption), 733954 drivers and 37978 relevant MVAs, 99989 hospital episodes	Alcohol consumption, MVA, hospital admissions (inpatient episodes involving alcohol intoxication or poisoning, MVA, motorcycle accidents, other external causes)	National household survey (HILDA), NSW Centre for Road Safety, National Hospital Morbidity Database
Yoruk and Ertan Yoruk 2011	Approx. 9000	Alcohol, cigarette, and marijuana consumption	NLSY97 longitudinal survey

The estimation of the effect of an intervention (or exposure) on an outcome within an RD design is achieved through regression analysis, for which the underlying functional form is typically unknown, making the estimate sensitive to model specification. There is no single or simple approach to identifying the best estimate or eliminating inappropriate specifications, so testing and reporting multiple specifications is standard practice in RD; relying only on one specification is not recommended.(Lee and Lemieux, 2010) Table 4.3 reports findings on how the studies conducted and reported modelling of the relationship between the forcing variable and outcomes. No studies reported the existence of a protocol or a priori method of model specification and selection.

Table 4.3. Characteristics of statistical analyses presented in RD studies of MLDA legislation

Study	No. models reported	Variables in model	Subgroups analysed	Model selection method	Description of preferred model
Boes and Stillman 2013 (date-based)	6	1/6 models adjusted for gender, ethnicity (admissions) or vehicle type (MVA), month of year, day of week, location (Table 4)	Age groups 15-17, 18-19, 20-21, 22-23 yos	Not reported; range of estimates reported rather than a preferred model	None chosen; range of estimates given
Boes and Stillman	6	1/6 models adjusted for gender, month of birth, vehicle type (MVA),	None	Not reported; approximate relative	None chosen; range of

Study	No. models reported	Variables in model	Subgroups analysed	Model selection method	Description of preferred model
2013 (age-based)		month of year, day of week, time of day, location (Table 8)		increase stated ("around 255")	estimates given in table
Callaghan et al. (2013a)	3	Birthday month, age	Gender (M, F)	Statistical significance of polynomial, described as 'the standard rationale for model selection'	Varies according to outcome and subgroup
Callaghan et al. 2013b	3	Birthday month, age	Gender (M, F)	Statistical significance of polynomial, described as 'the standard rationale for model selection'	Varies according to outcome and subgroup
Callaghan et al. 2014a	1	Birthday month, age	Gender (M, F)/MLDA 18 or 19	Polynomials for age tested and not significant, so only linear model presented	Linear with age interacted with MLDA
Callaghan et al. 2014b	1	Birthday week, age	Gender (M,F)	Polynomials for age tested and not significant, so only linear model presented	Linear with age interacted with MLDA
Carpenter and Dobkin 2009	4	1/4 models adjusted for birthday effect	None	p-value from Wald statistic for preferred parametric model, but all 4 presented (three parametric, one linear regression)	Quadratic polynomial in age interacted with MLDA dummy and adjusted for birthday month
Carpenter and Dobkin 2011	1	Birthday month, age	None	Only 1 reported	Quadratic polynomial in age fully interacted with MLDA dummy and adjusted for birthday month
Carpenter, Dobkin, and Warman (2014)	8	Age, birthday celebration effect. Appendix 9 additionally reports results for alcohol consumption adjusted for province,	Gender (M, F)	States "our preferred specification" without rationale, p. 13	Quadratic polynomial in age fully interacted with MLDA dummy and adjusted for

Study	No. models reported	Variables in model	Subgroups analysed	Model selection method	Description of preferred model
		year, month, and demographic variables			birthday month
Carpenter and Dobkin 2015	1	Birthday month, age	Gender (M, F)	Only 1 reported	Second-order quadratic polynomial in age
Conover and Scrimgeour 2013 (date-based)	5	Day of week and New Year holiday period dummies	Gender (M, F), age group (16-17, 18-19, 20-23)	BIC	"A linear model" but unclear which
Conover and Scrimgeour 2013 (age-based)	5	Year fixed effects	Gender (M, F)	Not reported; consistency of results across specifications observed and sensitivity discussed	None chosen; patterns described
Crost and Guerrero 2012	1	Age (donut approach to account for birthday month effect)	Gender (M, F)	Only 1 reported	Local linear with bandwidth of 3 years
Crost and Rees 2013	8	Age (donut approach to account for birthday month effect), household income, education, marital status, gender, race, student/employment status	None	Not reported; consistency of results across specifications observed and sensitivity discussed	None chosen; patterns described. Conclusion "we find no evidence that marijuana use changes at age 21"
Deza 2015	9 (three specifications with three models each)	Second specification includes birthday month effect, gender, race, educational enrolment, current or prior military service. Third spec includes fixed effects for time-invariant omitted variables	None	Not reported; consistency of results across specifications observed	None chosen; approximate values given which are not identical to model estimates
Ertan Yoruk and Yoruk 2012	7 (two parametric in Table 3 with two additional parametric and three non-parametric in Table 4)	Age, birthday celebration effect, household income, education, marital status, gender, race, student/employment status	None	Not reported; lack of statistical significance emphasised	None chosen; emphasis on lack of statistically significant effects
Ertan Yoruk and Yoruk 2015	6 (adjusted and unadjusted quadratic, cubic and quartic)	Age; 3/6 models adjusted for birthday celebration effect, household income, education, marital status, gender, race,	7 subgroup analyses plus 3 analyses of bandwidth sensitivity, all	Not reported	None chosen; emphasis on lack of statistically significant effects

Study	No. models reported	Variables in model	Subgroups analysed	Model selection method	Description of preferred model
		student/employment status	performed on quartic polynomial of age adjusted for observed control variables		
Lindo, Siminski and Yerokhin 2014	1 (with range of different bandwidths)	Age (donut approach to account for birthday month effect)	Gender (M, F)	IK optimal bandwidth selection procedure	Local linear with IK optimal bandwidth
Yoruk and Ertan Yoruk 2011	9 (four parametric and five non-parametric)	Age, birthday celebration effect, household income, education, marital status, gender, race, student/employment status	None	Not reported; consistency of results across specifications observed and sensitivity discussed	None chosen; range of estimates given

Reported model specifications varied across studies. Of the 17 studies, six reported only one specification (Callaghan 2014a and 2014b; Carpenter and Dobkin 2011 and 2015; Crost and Guerrero 2012; and Lindo et al. 2014), although one of these studies (Lindo et al. 2014) also reported results for a range of different bandwidths, and two of the studies (Callaghan et al. 2014 a and 2014b) tested other specifications but reported only a linear model after polynomials were tested and found not statistically significant. The other eleven studies reported between 3 and 9 specifications (median 6). Roughly half of the studies (9/17; 52.9%) additionally reported subgroup analyses by gender. All studies reported models that were adjusted for the forcing variable and for a 'birthday celebration effect', following Carpenter and Dobkin's original MLDA RD study (2009) which identified an immediate and short-term increase in mortality on the date of, or the day after, the 21st birthday.

Reporting of model selection methods and of preferred effect estimates varied considerably across studies. Of the eleven studies that reported estimates from more than one specification, one (Conover and Scrimgeour, 2013) reported using the Bayesian Information Criterion to select a preferred model for the age-based RD, although it was unclear which of the five models presented was selected and thereby represented the preferred estimate. Three studies (Callaghan 2013a and

2013b; Carpenter and Dobkin 2009) reported using statistical significance of the polynomial term for age as the rationale for model selection. One study (Carpenter 2014) identified a preferred specification but offered no rationale for the choice. The remaining six studies did not report any model selection method and did not report a preferred estimate, reporting instead a range of estimates and an approximation (Boes and Stillman, 2013), a lack of statistically significant estimates (Ertan Yoruk and Yoruk, 2012), or comments on the consistency of results across specifications (Conover and Scrimgeour, 2013; Crost and Rees 2013; Deza 2015; Yoruk and Ertan Yoruk, 2011).

4.5.4 Estimates of effects of MLDA

4.5.4.1 Mortality

Four studies reported estimates of the effect of reaching the MLDA on mortality. These studies used government death records from Canada (Callaghan et al. 2014a; Carpenter et al. 2014) or the United States (Carpenter and Dobkin 2009 and 2011). Estimates of effects on mortality due to all causes, MVA, and suicide were extracted and are reported in Table 4.4.

Table 4.4 Estimates of effect of minimum legal drinking age legislation on mortality

Study	All-cause mortality	MVA mortality	Suicide mortality
Callaghan et al. 2014a	Male MLDA 19 25.79 [SD 8.24] additional deaths or 7.2% increase, p=0.003; female MLDA 19 -0.21 [5.24] or -0.2% decrease, p=0.968	Male MLDA 19 22.05 [SD 5.50] additional deaths or 15.3% increase, p<0.001; male MLDA 18 12.7% increase, p<0.05. Female MLDA 19 2.09 [2.98] or 4.8% increase, p>0.05; female MLDA 18 13.6%, p>0.05 (figure 5)	NA
Carpenter and Dobkin 2009	8.7% increase (table 4)	14.3% increase (table 5)	15.4% increase (table 5)
Carpenter and Dobkin 2011	8.06 [SE 2.17] additional deaths per 100,000 person-years or an 8.7% increase, p<0.01; analysis by gender finds statistically significant effect for men only	3.65 [SE 1.25] additional deaths per 100,000 person-years or a 12.2% increase, p<0.01	2.37 [SE 0.76] additional deaths per 100,000 person-years or a 20.3% increase, p<0.01
Carpenter et al. 2014	4.10 [SE 2.76] additional deaths per 100,000 person-years or a 5.8% increase, p>0.05;	4.78 [SE 1.56] additional deaths per 100,000 person-years or a 17% increase, p<0.05; higher for males (7.32) than for females (2.12)	NA

	higher for males (6.91) than females (1.14) but neither increase is statistically significant, nor in the difference in effect by gender Also reported in Appendix 27: Provincial MLDA 19 0.25 [SE 2.73] increase in mortality rates; provincial MLDA 18 10.41 [SE 4.02] increase in mortality rates	and only statistically significant for men. Also reported in Appendix 27: Provincial MLDA 19 5.28 [SE 1.83] increase in mortality rates (19.8% relative increase, calculated); provincial MLDA 18 3.97 [SE 2.25] increase in mortality rates (12.8% relative increase, calculated)	

The two Canadian studies both found an increase in all-cause mortality for males; in Callaghan et al. (2014a) this relative increase of 7.2% in provinces with an MLDA of 19 (p=0.003) and 14.2% in provinces with an MLDA of 18 (p=0.002) was statistically significant, but in Carpenter et al. (2014) the increase (of 6.91 additional deaths per 100,000 person-years) was not statistically significant. Neither study found a statistically significant change in mortality for females. Both studies used Statistics Canada mortality data, although Carpenter et al. analysed the period 1980-2008 and Callaghan et al. analysed 1980-2009. Carpenter et al. analysed average mortality rates for each age-in-months, whereas Callaghan et al. analysed mortality counts for each age-in-months category. Carpenter et al. defined the birthday celebration effect as the birthday or week immediately after, whereas Callaghan et al. defined it as the birthday month. Both studies include individuals within two years of either side of the MLDA cutoff (48 age-in-month periods). Callaghan et al.'s model is linear whereas Carpenter et al. used a second order polynomial; both models are adjusted for interaction with the MLDA. A final difference between the studies is that Callaghan et al. conduct separate analyses for provinces with MLDAs of 18 and 19, respectively, while also stratifying by gender; Carpenter et al. combine all provinces in a single analysis, arguing that "separate analyses by provincial MLDA are not informative because the vast majority of the Canadian population resides in provinces with an MLDA of 19" (p. 4). (A footnote (p. 23) points to the presentation of further results in Appendix 27, where Carpenter et al. present mortality estimates stratified by provincial MLDA. Unfortunately, these are not stratified by gender.

The results of the two Canadian studies showed greater similarity when evaluating MVA mortality. Both found statistically significant increases in MVA

mortality at the MLDA for males, but not for females. The point estimates of the increase are similar when comparing the results for males in both provincial MLDA categories in Callaghan et al. with the results stratified by province (but not gender) in Carpenter et al. (see column 3 of table 4).

The two studies of American data, both by Carpenter and Dobkin (2009 and 2011), report the same statistically significant ($p<0.01$) estimated increase in all-cause mortality of 8.7%, although when stratified by gender the increase is statistically significant for males but not for females (web appendix O, 2009). The two papers give different estimates for effects on mortality from MVA or suicide but all estimates are increases and range from 12.2% to 20.3%. It is unclear why the estimates differ between the two papers when they use very similar analytical approaches and data (with one extra year of observations used for the 2009 paper).

4.5.4.2 Hospitalisation

Seven RD analyses in six studies examined changes in hospital admissions at the MLDA (Boes and Stillman, 2013; Callaghan et al, 2013a and 2013b; Carpenter and Dobkin, 2015; Conover and Scrimgeour, 2013, both age- and date-based RD designs; and Lindo et al., 2014). These studies cover all four countries represented in this review. All six studies used large administrative datasets and ICD-9 or ICD-10 codes to identify admissions related to alcohol, injury, or MVA, although the specific codes used and the composite outcomes created vary across studies. Table 4.5 presents the estimates of effect and definition used for alcohol-related hospital admissions.

Table 4.5 Estimates of effect of minimum legal drinking age legislation on alcohol-related hospital admissions

Study	Alcohol-related hospital admissions	ARHA definition
Boes and Stillman 2013 (date-based)	Lowering of MLDA from 20 to 18 in 1999 led to small absolute increases in ARHA for 15-21 year olds (24 point estimates presented "which range from about 0.3 to 0.4 additional admissions per 10,000 population for the 15-19 year-olds (significant at the 1% level) and 0.2 additional admissions per 10,000 population for the 20-21 year-olds (significant at the 5% level)". In relative terms ARHA "almost" doubled.	Admissions for alcohol use disorder, alcohol intoxication, and alcohol dependence as per ICD-9 codes

Callaghan et al. (2013a)	At MLDA there are statistically significant increases in admissions for alcohol use disorders/poisoning (M and F), self-inflicted injuries (total), MVA (M), and external injuries (M), but not assault and not for F other than alcohol use disorders/poisoning	Alcohol use disorders/poisoning (composite outcome), self-inflicted injuries, assault, MVA, and external injuries as per ICD-9 or 10
Callaghan et al. 2013b	(Looking at best-fit models only) No significant effect on suicide broadly defined, alcohol-use disorders in females, MVA; significant increase in alcohol-use disorders (male and all), assaults, alcohol-related suicide (female and all, but total events =11), male external injuries	Alcohol use disorders, suicides related to alcohol, suicides broadly defined, assault, MVA, and external injuries as per ICD-10
Carpenter and Dobkin 2015	At MLDA there are statistically significant increases in all ED visits (71.3 per 10,000 person-years) and hospital admissions (8.4 per 10,000 person-years). These are the highest estimates of effect for any of the outcomes/subgroups. If look only at E/D or hospital admission for alcohol intoxication, effects are much smaller (but still statistically significant)	All admissions/visits excluding pregnancy and for the following causes: alcohol intoxication, alcohol or injury (composite), accidental injury, self-inflicted injury, injury inflicted by other, as per ICD-9
Conover and Scrimgeour 2013 (date-based)	Following policy change (MLDA lowered to 18) there was an increase in hospitalizations of 53.663% (SE 24.605, $p<0.05$) for males and 4.673% (SE 14.153; NS) for females (linear estimate using one year's data)	Admissions coded with ICD-9 or -10 with mention of alcohol
Conover and Scrimgeour 2013 (age-based)	At MLDA hospitalizations increase by 19.395% (SE 8.452) for males ($p<0.05$) and decrease for females by -1.959% (SE 6.374) (linear estimate using one year's data)	Admissions coded with ICD-9 or -10 with mention of alcohol
Lindo, Siminski and Yerokhin 2014	At MLDA there is a statistically significant increase in hospital admissions for alcohol intoxication or poisoning of approximately 4 episodes per 10,000 person years or a 30% relative increase (similar magnitude for males and females), and approximately 7 episodes per 10,000 person years for assault (greater for males than females – approximately double). No evidence of discontinuity in admission for drivers injured in MVAs.	Alcohol intoxication or poisoning (ICD-10 alcohol use disorder or toxic effect of alcohol), assaults, transport accidents

Despite these similarities among studies, relatively clear definitions of outcomes, and use of large national datasets, it was challenging to extract data and summarise the findings of these studies. Table 5 reports the results. Boes and Stillman (2013), Callaghan et al. (2013 a and b), and Conover and Scrimgeour (2013) all report estimates from multiple model specifications; Callaghan et al. and Conover and Scrimgeour provide estimates only by gender. Estimates of effect were expressed as number of additional admissions per 10,000 population (Boes and Stillman, 2013), rates per 1,000 hospital events (Callaghan et al., 2013b), admissions per 10,000 person-years observed (Carpenter and Dobkin, 2015; Lindo et al., 2014), or relative increases in admissions (Conover and Scrimgeour, 2013; Lindo et al., 2014). Although

absolute numbers of events observed were sometimes reported, denominators generally were not. Denominators frequently were estimates based on government census data.

4.5.4.3 Motor vehicle accidents

Four RD analyses in three studies examined changes in MVAs at the MLDA (Boes and Stillman, 2013, both age- and date-based RD designs; Callaghan et al., 2014b; and Lindo et al., 2014), using data from three of the four countries represented in this review. All three studies used government data on MVAs.

The inconsistent findings among these analyses are described in table 6. Despite conducting numerous analyses using different subgroups of MVAs, Lindo et al. found no evidence of any discontinuity at the MLDA. Boes and Stillman identified a significant effect in their age-based approach, but not in their date-based RD analysis. Callaghan et al. found statistically significant increases in all MVA types and for almost all subgroup analyses, despite the fact that the study hypothesised that any effect of MLDA would only be seen in nighttime and single-vehicle nighttime crashes (their proxy outcomes for MVA involving alcohol). Table 4.6 presents the estimates of effect and definition used for alcohol-related hospital admissions.

Table 4.6 Estimates of effect of minimum legal drinking age legislation on motor vehicle accidents

Study	Motor Vehicle Accidents
Boes and Stillman 2013 (date-based)	(date-based discontinuity) "Overall…the reduction in the MLDA had no immediate impact" (p. 14 and table 5)
Boes and Stillman 2013 (age-based)	(age-based discontinuity) "Taken at face value, these results indicate that having the MLDA at 18 increases alcohol-related vehicular accidents by around 25%" (p.18)
Callaghan et al. 2014b	Statistically significant increases in total MVAs (all, daytime, nighttime, single-vehicle nighttime) and MVAs for men (all types except single-vehicle nighttime), but not for women (although all MVAs just reached statistical significance, p=0.473) (tables 1 and 2. Estimates of the increase ranged from 4.2% (women, daytime) to 16.3% (women, nighttime).
Lindo, Siminski and Yerokhin 2014	“Consistently, we found no evidence of discontinuities” (p. 18)

Boes and Stillman (2013) provide a unique insight by analysing the same data using the discontinuities created by two different natural experiments, the age threshold and the date that the age threshold was lowered in New Zealand. Unlike the data used in the other two studies, the New Zealand data did record whether the accidents were deemed by the police to be alcohol-related (p. 13). The date-based RD found that alcohol-related MVA did not increase following the change in MLDA legislation and the authors further confirm this finding through sensitivity checks and by estimating a difference-in-difference model (p. 14). An age-based RD of MVA during the period before the change in legislation, when the MLDA was 20, also found "little evidence" of a discontinuity. However, repeating the age-based RD for the period post-legislative change, Boes and Stillman found an increase of approximately 0.08 alcohol-related MVA per 10,000 population at the threshold (p. 17) or an approximate relative increase of 25%.

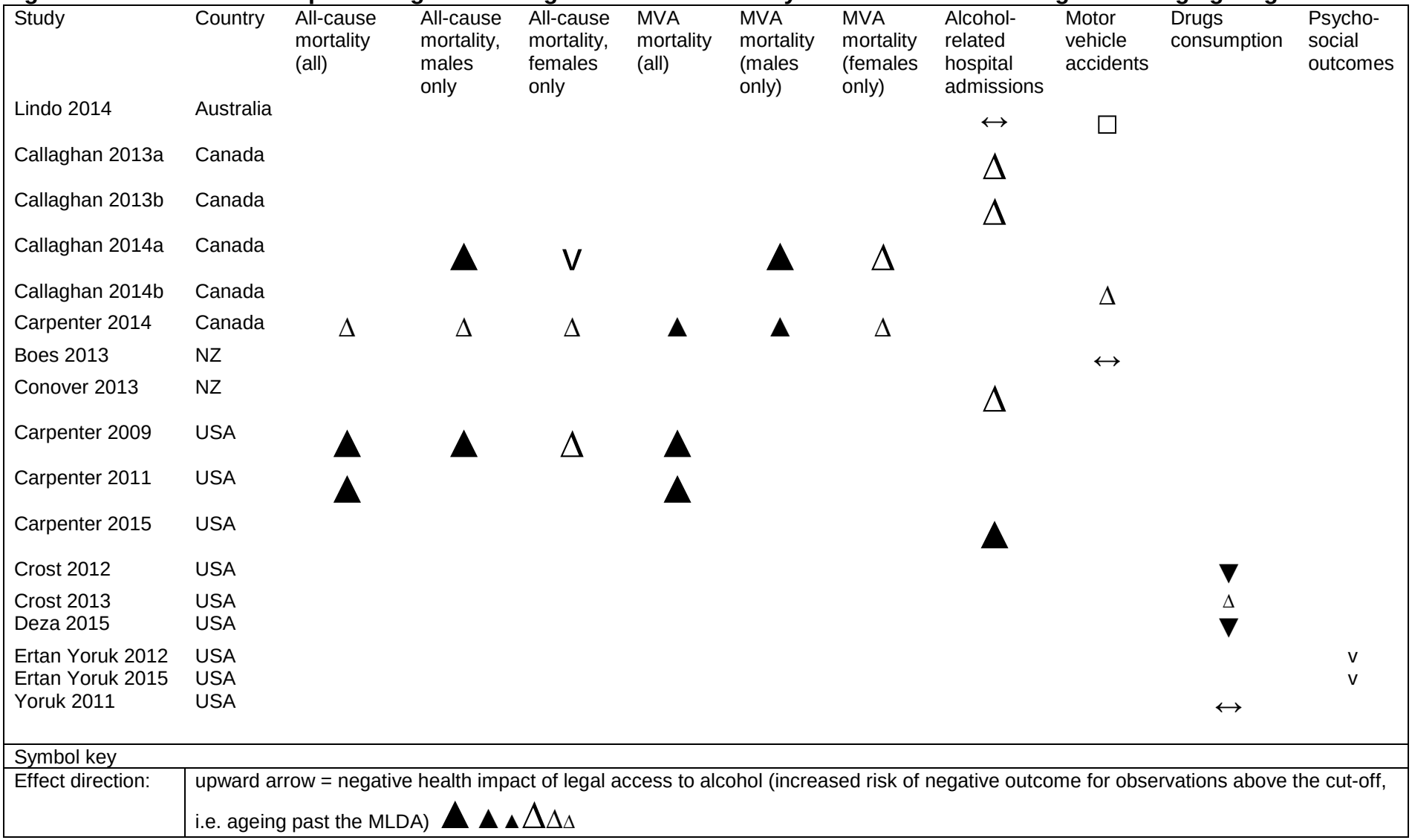
Boes and Stillman offer a three-pronged explanation of this inconsistency. First, they argue that the RD design is a LATE estimator which only identifies an effect on individuals whose behaviour is changed by the MLDA, meaning that those ageing past the threshold will be inexperienced drinkers more likely to experience a negative impact of increased access to alcohol. Second, they note that the distribution of MVA by age is an inverse U-shape and it is this nonlinear distribution that affects the RD estimate for the younger age group. Third, they offer contextual information suggesting that enforcement of MLDA legislation was lax prior to 1999 but increased following the law change. They conclude that their results "provide strong evidence that an age-based RDD is likely to give misleading evidence on the average impact of changing a MLDA, which is the policy relevant question" (p. 18).

4.5.4.4 Effect direction plot

The effect direction plot for this review (Figure 4.1) was designed to summarise the findings for all of the above outcomes in one table with graphical representation of effect estimates. Studies have been grouped by country of MLDA to aid comprehension of the body of evidence and judgment of any similarities and differences in effects. The outcome groupings 'drugs consumption' and 'psychosocial outcomes' have been added so that findings from all included studies can be represented in the plot. Drugs consumption

includes marijuana (Crost and Guerrero 2012; Crost and Rees 2013; Yoruk 2011) and ‘hard’ drugs such as cocaine (Deza 2015). Psychosocial outcomes include various psychological wellbeing measures (Ertan Yoruk 2012) and sexual behaviours (Ertan Yoruk 2015).

Figure 4.1. Effect direction plot for age-based regression discontinuity studies of minimum legal drinking age legislation



	downward arrow = positive health impact of legal access to alcohol (decreased risk of negative outcome for observations above the cut-off, i.e. ageing past the MLDA) ▼ ▼ ▾ V v v
	sideways arrow=mixed effects or conflicting findings ↔
	square: consistent evidence of no effect, i.e. no discontinuity, i.e. zero health impact □ □ ▫
Sample size:	size of arrow: large arrow > 100,000; medium arrow > 10,000 but <100,000; small arrow <10,000 records or participants in dataset or sample
Statistical significance:	Black arrow p<0.05; white arrow p>0.05
Synthesis of multiple model specifications within same outcome category	
1. Where multiple specifications all report effect in same direction and with same level of statistical significance, report accordingly.	
2. Where direction of effect varies across specifications, report direction of effect and statistical significance where at least 2/3 of specifications report same direction and similar statistical significance.	
If less than 2/3 of specifications report same direction of effect, then report no clear effect / conflicting findings (sideways arrow)	
3. Where direction of effect is similar across specifications but statistical significance varies:	
	If direction of effect is similar in at least 2/3 of model specifications AND at least 2/3 of specifications are statistically significant, report as statistically significant (black arrow).
	If direction of effect is similar in at least 2/3 of model specifications AND less than 2/3 of specifications are statistically significant, report as not statistically significant (gray arrow).

Assumptions: Boes and Stillman: large sample size based on 2001 New Zealand census (n=270,456 people aged 15-19). Source: Statistics New Zealand. Age Group and Sex, for the Census Night Population Count, 1991, 1996 and 2001. Available from: https://www.stats.govt.nz/tools/nz-dot-stat [accessed 19 June 2018]. Callaghan 2013b: large sample size based on reported number of hospital admissions per month (approximately 40,000) for five years of data. Callaghan et al. 2014b: medium sample size based on reported number of MVA involving at least one 18-year-old driver (n=70,585). Carpenter and Dobkin 2009 and 2011: large sample size representing all deaths in United States over an eight-year period (number not reported).

Reading down each column, the effect direction plot can be interpreted as follows:

- Evidence from four studies conducted in Canada and the United States suggests that mortality (from all causes and from MVA) increases at the MLDA and that this effect is statistically significant in males but not females.

- Evidence from five studies conducted in Australia, Canada, New Zealand, and the United States suggests that alcohol-related hospital admissions increase at the MLDA, but that this effect is probably not statistically significant and is not consistent across settings.

- Evidence is inconsistent from three studies conducted in Australia, Canada, and New Zealand on the effect on MVA at the MLDA.

- Evidence is inconsistent from four studies conducted in the United States on drugs consumption. However, the two larger (and higher-quality) studies suggest a positive health effect, i.e. reduced drugs consumption.

- Evidence from two small studies conducted in the United States suggests a positive but not statistically significant effect on psychosocial outcomes at the MLDA.

4.6 Discussion

This chapter has examined a subset of studies from a systematic review of RD in health (chapter 3), as per the review protocol which specified that further analyses would be undertaken if multiple studies were identified on the same topic. Seventeen studies investigated the health effects of MLDA legislation. This chapter reported the characteristics of those studies, quality assessments of the studies based on the WWC RD standards, and a synthesis of findings on the effects of MLDA legislation (or more precisely, the effect of ageing past the threshold for legal drinking) on mortality, hospital admissions, MVA, drugs consumption, and psychosocial outcomes.

4.6.1 Interpretation of results: evidence on effectiveness of MLDA

This review shows that RD studies of the MLDA provide consistent evidence that mortality, fatal MVA, and alcohol-related hospital admissions increase when people age past the MLDA threshold, i.e. when age-based restrictions on alcohol are removed. This evidence suggests that MLDA legislation is effective in preventing alcohol-related harms in people younger than the age cut-off. The policy implication is that lowering (or removing) the age limit would expose people to these risks of harm at an earlier age and thereby increase social costs (through more life-years lost or through earlier onset of disabling conditions). These findings are consistent with previous systematic reviews that did not include RD studies (Shults et al., 2001, Wagenaar and Toomey, 2002).

The effectiveness of MLDA was not consistent across all outcomes (although it must be kept in mind that the number of studies for each outcome is low). Contrary to previous reviews, the narrative synthesis and effect direction plot in this review suggest that there is inconsistent evidence on the effect of the MLDA on MVA. Lindo et al. (2013) found no evidence of a change in MVA at the threshold and argue that this finding reflects the "relative seriousness" with which New South Wales enforces and penalises drink-driving (p. 21), such that MLDA does not perceptibly change driver behaviour. Callaghan et al. (2014b), although emphasising statistically significant findings, especially in men, in fact reported a mixture of positive and negative findings with wide confidence intervals, and found a statistically significant increase in daylight MVA, an outcome which they suggested should not be predominantly affected by alcohol.

Although the number of RD studies of this outcome is small, it is worth giving some consideration to possible reasons for differences between the findings of these studies and previous MLDA reviews, such as Shults et al. who found strong evidence for the effectiveness of MLDA in reducing alcohol-related MVA. Granted, these RD studies were conducted in countries other than the United States, and although MLDA operates on principles that are cross-cultural, factors such as enforcement and social norms may lead to different effects of MLDA in different countries. However, the impact of different study designs should also

be considered. Is it possible that previous reviews of observational studies mistook association for causation (McCartt et al., 2010)? Or could RD studies provide inaccurate and misleading results?

To explore these questions, it is helpful to look closely at an example from New Zealand. Boes and Stillman (2013) also reported mixed findings on MVA, with implications not just for the effectiveness of MLDA but also for the interpretation of age-based versus date-based RD. Their date-based RD found that after the lowering of the MLDA December 1999, MVA decreased for all four age groups examined in the majority of model specifications; in the 18-19 year old age group which newly had legal access to alcohol in this situation, all specifications found a decrease in MVA, although only two of six estimates were statistically significant. Perhaps because the majority of estimates for all age groups were not statistically significant, they concluded that the 1999 reduction in the MLDA "had no immediate impact" on MVA (p. 14).

Their age-based RD, on the other hand, found a small but statistically significant increase in MVA at the age threshold of 18 in the period 2000-2007, i.e. for drivers with newly-acquired legal access to alcohol. This increase was consistent in direction and significance across six specifications. The point estimate of the increase ranges between 0.065 to 0.099 alcohol-related MVA per 10,000 population. Boes and Stillman interpret the disagreement between age-based and date-based RD as follows:

> "given that the results in the previous section, which identify the impact of the MLDA using the policy change itself, show no impact of moving the MLDA to 18, we believe the results here provide strong evidence that an age-based RDD is likely to give misleading evidence on the average impact of changing a MLDA, which is the policy relevant question." (p. 18)

Before accepting that the age-based RD design is misleading and inaccurate compared to a date-based design, it is necessary to closely consider the natural experiment being assessed in each design. Boes and Stillman interpret the age-based RD to estimate the effect of the change of legislation. However, using age as the forcing variable means that the RD design is estimating the effect of reaching the age threshold and thereby gaining legal access to alcohol – not the effect of introducing new legislation. The date-based RD investigates that

change, but it also assumes that no other significant changes occur at that date. In fact, the 1999 SLA included several changes affecting access to alcohol in addition to the MLDA change. Thus, it is entirely plausible that different effects could be produced by the two different RDDs, because they are evaluating two different exposures and two different natural experiments.

That stated, it is also worth considering the evidence from controlled before-and-after studies of the effect on MVA of lowering the MLDA from 20 to 18 in New Zealand. These studies found statistically significant increases in MVA in the under-20 age group following the law change (Huckle and Parker, 2014, Kypri et al., 2017, Kypri et al., 2006). Perhaps it is the date-based RD rather than the age-based design that produces misleading evidence on the effects of the legislative change. How estimates of effect may vary according to the study design of natural experiments with date- or time-based cutoffs would appear to warrant further investigation.

4.6.2 Implications for alcohol policy

The evidence from RD studies on mortality and alcohol-related hospital admissions supports the place of MLDA in alcohol policy as a public health intervention that reduces a range of alcohol-related harms and societal costs. The findings support the WHO Global Strategy recommendation to establish a minimum age for purchase and consumption as an effective policy option for reducing availability of alcohol (World Health Organization, 2010). The inconsistent evidence on MVA may be interpreted to mean that other policy options should be explored if reduction of MVA is the policy objective. The protective effect of MLDA appears to be larger, and the supporting evidence stronger, for men compared to women. Policymakers should note gender differences in the evidence base for reduction of alcohol-related harms and ensure that related policy frameworks and strategies include a mixture of interventions that will, in toto, be effective for both groups so as not to inadvertently increase gender inequality.

Despite, or perhaps because of, the apparent lack of interest in MLDA as an alcohol policy option in the UK, policymakers in this country should seriously consider the potential reductions in alcohol-related harms and related societal

costs that could be achieved by setting, and enforcing, an appropriate legal drinking age. It seems shocking to have to say, given the evidence of binge drinking, alcohol-related hospital admissions, and related violent offending among UK adolescents (Healey et al., 2014), that the age of five should no longer be considered an appropriate MLDA in this country (Gerard, 2007).

4.6.3 Implications for research

This section discusses the implications of the findings for future research in relation to three topics. First, areas for development of systematic review methods are considered. Second, implications for the design and reporting of RD studies are discussed. Finally, some possibilities for further research into MLDA are described, including extension of the present review and further potential applications of RD to evaluate MLDA in different settings.

4.6.3.1 Methods for systematic review of natural experiments

This review demonstrates the potential for RD studies to be incorporated into evidence syntheses and to inform public health policy. Systematic review methods require some adaptation to achieve this end. The PICO (Population, Intervention, Comparison, Outcome) model may be a loose fit for natural experiments; studies of MLDA legislation actually represented two different natural experiments, effects of ageing past the threshold (removal of age-based restriction) and effects of changing the legislation. Critical appraisal tools and synthesis methods may require adaptation to fit the different study design and different conventions of reporting in disciplines other than health. There is a need to develop methods for synthesising effect estimates from multiple models; the effect direction plot is helpful, but requires further testing and application. Poor statistical reporting means that meta-analysis of such studies is likely to require author contact, which takes up time and resources. Reviews of natural experiments also would likely benefit from synthesising information about the context and implementation of the intervention, not just the results. This would also require additional time and resources, as well as considerations at the protocol and data extraction stages.

The reporting of data in these studies represented a significant challenge for the review. No studies clearly reported the number of events and observations (numerators and denominators) involved in each analysis. Most of the studies presented multiple model specifications, sensitivity analyses, and estimates of effect, with no clear rationale for choosing among them. All but one study lacked a primary outcome. Confidence intervals were generally absent. One study (Lindo et al. 2014) presented all of its RD estimates as a series of visual plots without any tabular presentation of data. Essential data sometimes appeared in footnotes or online appendices. These challenges were exacerbated by different conventions in reporting in economics compared to health research; however, these challenges were encountered consistently in the included studies, regardless of whether they were published in economics, health economics, or health research journals.

Given the data issues, the potential for meta-analysis was severely limited. The effect direction plot represented a useful method of visualising the results. In this review, construction of the effect direction plot was rendered more difficult by non-reporting of included numbers in each study. I solved this problem by creating categories that reflected the relative size of the included studies, which were based on large surveys, larger administrative datasets, and very large datasets such as census data. However, I had not anticipated this during initial data extraction, which meant that it was necessary to revisit all the studies in order to 'code' them for the effect direction plot.

The effect direction plot was designed to produce a synthesis that combined multiple related outcome measures into a single domain, which simplified reporting and aided interpretation. It solved a problem of interpreting study findings when results may vary across multiple related outcomes. This situation applied to many MLDA studies, which used multiple measures of mortality, hospital admissions, or MVA. However, an even bigger challenge was interpreting the results across multiple model specifications, so I decided to extend the principle to the synthesis of findings across outcomes and specifications. This, too, necessitated extracting additional data and sometimes needing 2X2 tables to apply the decision rules about direction and statistical significance in order to code the outcome for the plot. This process would become easier with greater

familiarity, and easier still if anticipated earlier in the review and incorporated into the data extraction and synthesis plan. Overall, the effect direction plot is a good tool for summarising and visualising the results of natural experimental studies that involve multiple models and multiple outcomes.

4.6.3.2 Design and reporting of natural experiments using RD

The critical appraisal and data extraction performed in this review point to several areas for improvement in the conduct and reporting of RD studies. Many studies lacked basic details such as numbers included in analyses, information about attrition, and uncertainty of effect estimates. No studies were protocol-driven, only one specified a primary outcome, and model selection methods were often unclear. A standard for reporting could help to improve RD studies, as CONSORT did for RCTs, which would have the additional benefit of making systematic reviews of such studies both easier to perform and more informative.

Contextual information about the natural experiment under investigation is necessary in order to understand the hypothesis being tested and to assess the validity of the RD design. Most of the MLDA studies provided little narrative justification for the validity of the RD design and little or no statistical investigation of related assumptions, perhaps because the legislation seemed relatively straightforward and the assumption that age cannot be manipulated seemed reasonable, or perhaps too obvious to mention. However, contextual information is important for understanding differences between studies and explaining inconsistent results. Moreover, study quality is improved and the strength of the overall evidence base may be increased if design assumptions, such as smoothness of the forcing variable, are investigated. Improving the standard of design and reporting is important if natural experimental studies are to fulfil their potential to contribute to the public health evidence base.

4.6.3.3 Further MLDA research

The present review could usefully be expanded in at least three ways. First, alcohol consumption could be included as an outcome and a meta-analysis might be possible. Second, it would be informative to investigate the context of MLDA in these studies, obtaining information from the study reports and from

additional sources in order to better understand the natural experiments. Conclusions could then be drawn about reasons for differences between settings and what information would be useful to include in reports of natural experiments. Third, in order to better understand the place of RD in a larger body of evidence and to understand the sensitivity of results to study design, it would be useful to conduct a larger review of MLDA including study designs such as controlled before and after, difference-in-differences, and interrupted time series as well as RD. A review of MLDA evidence of such scope has not been reported since 2002. In addition to providing an updated synthesis, such a review would afford an opportunity to investigate and better understand whether and how different natural experimental designs differ in their estimates of effect.

This review has demonstrated the potential for Carpenter and Dobkin's original RD design to be replicated in different settings and to investigate various outcomes. With only 17 studies identified in four countries, there is potential to repeat the design in other countries that have an MLDA with appropriate enforcement and which is not confounded by other changes at the same age threshold, i.e., the cut-off has meaning and plausibly creates a discrete change in access to alcohol. One European candidate country for evaluation would be Iceland, with its MLDA of 20, strict drink-driving laws, and low perceived availability of alcohol to underage drinkers (The European School Survey Project on Alcohol and Other Drugs, 2015).

Further applications in new settings could also investigate extensions of the RD design. One possibility would be to investigate whether this design could evaluate policies like those in the UK and Germany, where different access to alcohol (types of alcohol and settings of purchase/consumption) becomes available at several different age thresholds, for example using the multiple cut-off RD design (Cattaneo et al., 2016). It would also be possible to use geographical boundaries as cut-offs to investigate the comparative effectiveness of different MLDAs where these vary between neighbouring countries. For example, Paraguay has an MLDA of 20 whereas all of its neighbours have an MLDA of 18.

4.6.3.4 Strengths and limitations

The outcomes selected for detailed data extraction (mortality, hospital admissions, and MVA) have undoubted policy relevance and are among the most costly of alcohol-related harms affecting young adults. Restricting the systematic review to health outcomes meant that other policy-relevant outcomes with high social costs, such as crime, have not been included. RD studies on this topic (published after the search cut-off date) exist (Callaghan et al., 2016a, Callaghan et al., 2016b, Carpenter and Dobkin, 2015a) and it can be argued that crime, particularly violent crime, is an outcome of interest to public health.

Alcohol consumption was an outcome measured by most of the seventeen studies. I did not include this outcome in data extraction or synthesis for two reasons. First, an increase in consumption of any desirable and plentiful commodity following a price decrease (removal of age restriction) can hardly come as a surprise (although a lack of discontinuity would be informative regarding compliance with and enforcement of MLDA laws). Second, in terms of causality I considered alcohol consumption an intermediary outcome whose sequelae, such as hospitalisation and mortality, were of greater policy relevance. For this reason I focused on outcomes that were further downstream.

The design and conduct of this systematic review has several strengths. With reference to AMSTAR-2 criteria for the quality of systematic reviews (Shea et al., 2017), the review has been well conducted as it was based on a pre-published protocol with explicit inclusion criteria and had a robust and reproducible literature search. Characteristics of included studies were described in detail, duplicate study selection was performed on a sample of studies, risk of bias was assessed independently by two reviewers, and implications of risk of bias were considered and discussed.

The AMSTAR-2 criteria also point to several limitations of the review. Duplicate extraction of data was not performed owing to resource limitations. This may be of particular concern given the complexity of these studies and their reporting. However, the data extraction in this review was exploratory, being a novel application of systematic review methods to this study type. As such, duplicate

extraction may have added another layer of difficulty and complexity, of unknown utility, to this activity, which furthermore might have hindered the production of any synthesis. Further extension, development, and validation of the methods tested here, from one reviewer to a review team and from one natural experiment design to others, could be pursued in future research.

The investigation of heterogeneity in this review was limited, partly by the non-quantitative nature of the synthesis, but also by the focus of the review on methods rather than study contexts. Some attention has been paid to heterogeneity created by study setting and by type of forcing variable, but further consideration could be given to other factors that might have influenced the varying findings. The statistical reporting of outcomes in these studies would not lend itself to meta-regression, however. What would be most useful would be a consideration of the details of the natural experiment itself: the contexts and mechanisms of MLDA. Extraction of such information from the studies, and collation of supporting information from external sources, was beyond the scope of this review.

An AMSTAR-2 appraisal would also point out that the review failed to consider the funding sources of these studies and to investigate publication bias. Study funding is potentially important as it would be in the interest of the alcohol industry to fund research that supported lowering the MLDA or that reported low risks of alcohol-related harms. Current methods of assessing publication bias would need to be adapted as they rely on consistent reporting of effect sizes across studies, which cannot at present be observed in the RD literature. However, in the present sample of studies both positive and negative as well as conflicting findings were identified within and across studies, grey literature was searched, and numerous unpublished studies were identified, all of which suggest that this review is at low of risk of bias in these domains.

A final limitation relates to the currency of the review. The last date searched was March 2015 and at least two new RD studies of MLDA and health outcomes have been published since then.(Callaghan et al., 2016c, Koppa, 2018) Callaghan et al. extend their study on Quebec (2014b) to six other provinces and the Northwest Territories of Canada, reporting effects above the MLDA on alcohol-related motor vehicle collisions and night-time motor vehicles collisions. Each

outcome is reported (where data are available) for males, females, and total population separately by province/territory, resulting in a mixture of positive and negative effects of varying statistical significance. Koppa (2018) examines data from California to investigate whether there is an increase in cases of gonorrhoea at the MLDA threshold and finds no evidence of an increase. Both of these studies would be useful to add to the effect direction plot, particularly given the small number of studies investigating MVAs and psychosocial outcomes (including sexual behaviours), but would not change the conclusions of the review or the implications for policy.

4.6.4 Contribution of this systematic review

This chapter presents the first systematic review of an intervention or policy effectiveness question restricted to RD studies; the first application of the effect direction plot to RD; and the first systematic synthesis of RD evidence on MLDA legislation. As such it represents a proof of concept for several points relevant to encouraging the creation and uptake of evidence from natural experiments, showing that it is possible to replicate a natural experiment design in different contexts with different data, synthesise such evidence, and thereby reduce the uncertainty associated with the findings of single studies. Furthermore, it demonstrates that policy-relevant conclusions can be drawn even though randomised trials are lacking, the reporting of RD studies poses challenges, and synthesis is complex.

The key contributions of this chapter are methodological, specifically knowledge about the application of RD designs and the incorporation of these studies in systematic reviews. Although Angrist and Pischke (2015) claimed that Carpenter and Dobkin's original MLDA study design was "made in RD heaven", this review has demonstrated some limitations in the MLDA studies when compared to standards for RD. Close examination of these studies also shows that careful consideration of the setup of the natural experiment and the implementation context of the legislation are necessary to determine exactly what hypothesis is being tested; studies that appear to have essentially the same RD design may in fact be answering quite different research questions. Data extraction for this systematic review further demonstrated that an apparently straightforward and intuitive design in fact poses considerable challenges for systematic review,

which could constitute a barrier for uptake of this evidence and for further implementation of the RD design. However, ultimately these challenges could be overcome with only minor adaptations of existing synthesis methods. This chapter, then, serves as a positive example of the potential for natural experiments generally and RD designs in particular to be incorporated into systematic reviews and to usefully inform the public health evidence base.

Although this systematic review was designed to contribute to methodological knowledge, the relevance to alcohol policy adds some further value to this work. As randomised trials of age-restricted access to alcohol or other unhealthy commodities are not likely to be feasible or acceptable to legislators, regression discontinuity designs are likely to represent the best available evidence on the effects of such legislative interventions. Despite the challenges posed by the evidence, it was possible, through narrative synthesis, to make clear statements on the evidence of effectiveness and the areas of uncertainty. This review supports conclusions about the effects of MLDA legislation on important and policy-relevant outcomes (mortality, MVA, alcohol-related hospital admissions, and drug use) which could be used to inform public health decision making and policy intended to prevent alcohol-related harms in young people. In particular, it shows that existing MLDA laws probably have an overall protective health effect on young people who are prevented legal access to alcohol, but that the marked reduction in MVA shown in earlier observational studies of MLDA is no longer evident. Three possible explanations present themselves. The effect of MLDA on MVA may have been confounded in the earlier observational studies, it may have been caused by intervention components or implementation factors other than the age-based purchase restriction, or it may have been an effect subject to fade-out over time.

4.7 Chapter summary

This chapter has reported a systematic review of 17 RD studies of minimum legal drinking age (MLDA) legislation in four countries. The review provides updated evidence of the effects of MLDA on policy-relevant outcomes including mortality, hospital admissions, and motor vehicle accidents. It is innovative in applying a visual synthesis method, the effect direction plot, to RD studies. Poor reporting in these studies was evidenced through the results of critical appraisal and

through difficulties and gaps in data extraction. In response to Angrist and Pischke's view that MLDA is a natural experiment "made in RD heaven", this review suggests that although there is good reason to hope, heaven on earth has yet to be attained.

The next chapter completes the reporting of the systematic review of RD studies in health that began in chapter 3 and continued in chapter 4. Chapter 5 describes how an approach to quality assessment of RD studies was developed for this thesis. Both the adaptation of the effect direction plot (reported in chapter 4) and the development of a critical appraisal checklist for RD (reported in chapter 5) constitute contributions made by this thesis to systematic review methodology.

5 Critical appraisal of regression discontinuity studies

5.1 Chapter overview

Critical appraisal is an essential component of systematic review. In the absence of tools to support the appraisal of natural experimental studies, the development of critical appraisal methods is necessary if systematic reviews are to make greater use of evidence using these designs and the value of natural experiments as evidence is to be appreciated. This chapter reports research undertaken to identify, test, and develop methods for the critical appraisal of regression discontinuity (RD) studies. A literature search identified one design-specific quality assessment tool, the What Works Clearinghouse Standards for RD. As this tool was developed to assess evidence for educational interventions prior to their implementation in schools, I tested its face validity and applicability on a sample of 17 RD studies in health incorporating assessments conducted by a second independent reviewer. Based on these results, I modified the standards to produce a 10-item checklist, RD-10. I then tested the usability and applicability of RD-10 on a sample of 13 RD studies. This assessment incorporated assessments conducted by three independent reviewers. Finally the checklist was applied in the systematic review of 181 RD studies reported in chapter 3. On the basis of these experiences, I suggest further refinements of the tool.

5.2 Aims

This chapter aims to:

1. test a published quality assessment tool for RD, the US Department of Education What Works Clearinghouse Standards for RD (WWC), using a sample of studies that evaluate a public health intervention (minimum legal drinking age legislation)

2. develop a critical appraisal checklist for RD that is applicable to health research and useable in systematic reviews

3. test the checklist on a further sample of studies before applying the checklist in a comprehensive review of RD studies in health.

5.3 Background

5.3.1 Introduction

Systematic review is an important method within the evidence-based paradigm because it can produce a trustworthy and comprehensive representation of available evidence, which then can be accessed in a single publication. By increasing the accessibility of the evidence and presenting conclusions based on its totality, systematic reviews can act as a facilitator of evidence-based decision-making (Petticrew and Roberts, 2006, pp. 11-12). However, given that reviews synthesise information from multiple individual studies, with inevitable loss of detail from the individual study reports, there is the potential for information from studies that have been poorly designed and conducted and/or studies at high risk of bias to be reproduced uncritically. A review could then unintentionally increase the dissemination of biased results and even lend credence to them, which could ultimately result in decisions being made based on flawed or erroneous conclusions, with potentially harmful effects.

In order to avoid such unintended consequences, textbooks, handbooks, the PRISMA standard, and the AMSTAR-2 checklist for systematic reviews all agree on the need for reviews to critically appraise included studies, report the appraisal findings in the results, and take risk of bias (RoB) into consideration in drawing conclusions (Egger, Smith, and Altman, 2001; Higgins and Green, 2011; Petticrew and Roberts, 2006; Shea et al., 2017). However, there is a lack of critical appraisal tools specific to natural experimental designs.

This section will provide the necessary context for the work I conducted to test and develop methods of critical appraisal of RD studies. The section covers the principles of critical appraisal including definitions of study quality; principles for evaluation and selection of tools; the availability of existing critical appraisal tools; and the rationale for testing and developing a new tool.

5.3.2 Principles of critical appraisal

The Dictionary of Epidemiology defines critical appraisal as "Application of rules of evidence to a study to assess the validity of the data, completeness of reporting, methods and procedures, significance of results, conclusions, compliance with ethical standards, etc." (Porta, 2014), a potentially broad field of enquiry. The object of critical appraisal is sometimes more simply described as 'study quality', but in defining this term it quickly becomes apparent why critical appraisal is subjective. Deeks et al. (2003) in their review of critical appraisal tools for NRS note that study quality is "a rather subjective concept, open to different interpretations depending on the reader" (p. 23) and cite the definition used by Moher et al. (1995) in their review of RCT appraisal tools: "the confidence that the trial design, conduct and analysis has minimised or avoided biases in its treatment comparisons". The Cochrane Handbook notes that assessment of study quality "suggests an investigation of the extent to which study authors conducted their research to the highest possible standards" (section 8.2.2) and sets out its reasons for focusing instead on internal validity and assessment of risk of bias within critical appraisal, not least of which is that a study conducted to the highest possible standards might still be at a very high risk of bias. Bias has the advantage of a less contentious definition than quality: "a systematic error, or deviation from the truth, in results or inferences" (section 8.2.1); yet the concepts of both 'truth' and 'risk' return us to an epistemological situation in which subjective opinion is highly operative.

Critical appraisal tools can help to reduce the influence of subjective opinion, assist the reviewer in investigating the many different aspects of a study in which risk of bias may operate, structure discussions between reviewers, and organise information about risk of bias for presentation in the findings of the review. The tools may be presented as scales, checklists, or domain-based evaluations (Higgins and Green, 2011). Scales assign points according to the presence or absence of study characteristics, resulting in a total score for quality. This approach is specifically discouraged in the Cochrane Handbook. Checklists aid in the identification and recording of relevant information. Domain-based risk of bias evaluation with 'signalling questions' is the approach taken by the Cochrane Risk of Bias tool for RCTs and the ROBINS-I tool for NRS. Deeks et al. (2003) note that the content of appraisal tools can be identified

through two approaches, ‘threats to validity’ as identified by Cook and Campbell in their work on quasi-experimental designs (Shadish, Cook and Campbell, 2002) or ‘methods-description’ in which the characteristics of the reported method are recorded.

The publication of the Cochrane RoB tool for RCTs laid out seven principles for assessing risk of bias which have been influential in shaping subsequent methodological research practice. Higgins et al. (2011) advised against quality scales in favour of a focus on internal validity, which should be assessed on the basis of the trial results and not quality of reporting or other aspects of trial conduct such as ethical approval or statistical power. They accepted that critical appraisal requires judgment while at the same time arguing that it should be based on the assessment of domains chosen for a combination of theoretical and empirical reasons. They asserted that judgments of high or low risk of bias need to be specific to the data and outcomes as represented in the review, which may differ from the risk of bias in the overall report of each individual study.

5.3.3 Evaluation and selection of critical appraisal tools

Somewhat surprisingly given the importance of critical appraisal within systematic review, there is no quality standard for critical appraisal tools. Perhaps the most relevant information comes from Viswanathan et al. (2017), who convened a working group to update AHRQ methodological guidance and produced 18 recommendations covering the planning and conduct of risk of bias assessment in systematic reviews. Given the plethora of tools available and noting the lack of any suitable universal tool, Viswanathan et al. provided principles for selection, arguing that reviewers should choose tools that:

- were specifically developed for use in systematic reviews
- are specific to each study design being assessed
- address domains of bias through specific items
- are at least based on theory, and preferably on empirical evidence of bias, or “have reasonable face validity”
- avoid numeric scores.

Further relevant considerations can be extracted from the criticisms Deeks et al. noted that are commonly made of critical appraisal tools: failure to provide a

rationale for appraisal criteria, inclusion of criteria of uncertain relevance to study quality, and neglect of the methods of scale development (p. 36). These methods should follow four steps: "preliminary conceptual decisions; item generation and assessment of face validity; field trials to assess frequency of endorsement, consistency and construct validity; and generation of a refined instrument" (Deeks et al., 2003, p. 36).

5.3.4 Availability of critical appraisal tools

The development of systematic review methods over time has seen a proliferation of critical appraisal tools. Systematic reviews of critical appraisal tools have identified 34 different tools for the appraisal of RCTs (Moher et al., 1995), 40 for clinical practice guidelines (Siering et al, 2013), and 194 for non-randomised studies (Deeks et al., 2003). The tools vary in content and complexity; the number of items in the tools identified by Moher et al. ranged from 3 to 57, for example, while in Deeks et al. the number of items was between 3 and 103 (2003, Appendix 3).

Critical appraisal of non-randomised studies remains an active area of methodological development. The Cochrane Risk of Bias Methods Group and Statistical Methods Group have recently produced the ROBINS-I ("Risk Of Bias In Non-randomised Studies – of Interventions") tool (Sterne et al., 2016), incorporating the domain-based approach familiar to users of the RoB tool for RCTs while also providing specific evaluation criteria (or "signalling questions") for cohort and case-control study designs. The tool has three sections: one for review protocol considerations such as the PICO of interest, one which asks the user to specify a "target trial" or hypothetical RCT that would answer the review question, and one that focuses on risk of bias of NRS. The seven RoB domains investigated are confounding, selection of participants into the study, classification of interventions, deviation from intended interventions, missing data, outcome measurement, and selective outcome reporting. ROBINS-I was developed through informal expert consensus and repeated rounds of revision. User feedback was obtained through telephone interviews and training workshops (number of participants not reported). The first training workshop involved application of the tool to six NRS.

Thomson et al. (2018) tested the applicability of ROBINS-I for the assessment of risk of bias in public health natural experiments. Although the tool was helpful in articulating risk of bias, many elements of the tool were difficult to apply to natural experimental studies, the tool required a high level of epidemiological expertise to interpret, the accompanying guidance did not address issues that arose relating to applicability to natural experiments, and agreement among reviewers was poor. While these experiences demonstrated that reporting quality of natural experiments needs to improve, the authors also concluded that revisions to ROBINS-I would be helpful to address the level of difficulty for users and problems with applicability to natural experiments in public health.

5.3.5 Rationale for testing and developing a tool for RD

As described in chapter 2, the study designs and methods for analysing natural experiments have largely been developed in disciplines other than health. As natural experimental methods remain less familiar to health researchers than randomised studies or the designs commonly used in epidemiology, critical appraisal tools for natural experiments are lacking in the evidence-based toolkit and no comprehensive effort has yet been reported to identify domains or criteria for risk of bias across natural experimental designs. Within the Cochrane Collaboration, work is underway to expand the ROBINS-I approach to encompass regression discontinuity designs and interrupted time series (personal communication). Until these tools are published and performance-tested, health researchers conducting systematic reviews that include natural experiments have the options of applying generic non-randomised study assessment tools, borrowing design-specific tools from other disciplines, adapting tools, or developing new tools. Indeed, even as the ROBINS-I approach expands, there may be an ongoing unmet need for tools that are more straightforward to use and that do not require a high level of specialist knowledge to apply, particularly given the added technical difficulties of appraisal of NRS and the increased level of subjective judgment required (Thomson et al., 2018; Waddington et al., 2017).

In keeping with the principles of evaluation and selection of tools described in section 5.3.3, I sought a design-specific critical appraisal tool with detailed criteria for investigating the internal validity of RD designs. This tool needed to

be applicable to RD studies in public health, which may be retrospective and based on population-level data, and suitable for use in a systematic review.

5.4 Methods

This section describes the methods used to develop and implement a design-specific approach to the critical appraisal of RD. The development involved four steps:

1. A literature search was undertaken to identify any existing appraisal tools.

2. A tool identified by the search, the What Works Clearinghouse Standards for RD (WWC), was tested on a sample of 17 papers (the MLDA studies reviewed in chapter 4).

3. A ten-item checklist was adapted from WWC and tested on a purposive sample of 13 papers.

4. The ten-item checklist was applied to all studies identified in the systematic review of RD in health (reported in chapter 3). Each of these steps is described in the sections below.

5.4.1 Literature search for existing tools

The systematic review of RD studies of health outcomes involved a comprehensive literature search using the terms “regression discontinuity” and “regression-discontinuity” as keywords or free text terms. This search (reported in section 3.3.2) was broad enough to capture any quality standards or appraisal tools available from the resources covered. I supplemented the electronic searches by hand-searching textbooks, methodological papers, and the websites of systematic review and guideline development organisations for potential quality assessment tools. Box 5.1 lists the resources that were hand-searched.

Box 5.1 Handsearching for RD critical appraisal tools

Textbooks and Handbooks:
Cochrane Handbook
Dunning, Natural Experiments in the Social Sciences
Shadish, Cook, and Campbell, Experimental and Quasi-Experimental Studies
Petticrew and Roberts, Systematic Reviews in the Social Sciences
Angrist and Pischke, Mostly Harmless Econometrics

Methodological Papers:
Deeks et al. 2003
Imbens and Lemieux 2007
Lee and Lemieux 2010

Websites:
NICE
SIGN
EQUATOR
Critical Appraisal Skills Programme (CASP)
AHRQ Effective Health Care Program
Cochrane Effective Practice and Organisation of Care (EPOC)

The only tool identified in the search was WWC (Schochet et al., 2010). This finding was confirmed by a recent review of appraisal tools for quasi-experimental designs; of the 14 tools identified in that review, only WWC addressed both study design and methods of analysis in RD (Waddington et al., 2017). That review concluded that current appraisal tools are inadequate for consistent and "appropriate" evaluation of quasi-experimental designs (p. 50).

5.4.2 Pilot of WWC Standards

As critical appraisal should be conducted by two reviewers, I worked with a second reviewer to pilot the only published RD-specific appraisal tool (WWC) on a sample of RD studies. Both reviewers had more than ten years of experience in conducting systematic reviews involving non-randomised studies. The purpose of the pilot was (1) to determine whether the selected (health-related, non-educational) studies reported the information necessary to make a judgment against the criteria, (2) to decide whether the criteria had face validity or apparent usefulness in investigating the quality of health studies, (3) to assess the feasibility of using the tool in terms of time requirements and difficulty of application, and (4) to determine informally whether interrater agreement was

satisfactory or whether extensive disagreement might indicate additional problems with feasibility and face validity.

The sample of RD studies assessed in the pilot consisted of 17 studies of minimum legal drinking age (MLDA) legislation. This sample was selected because it was the largest number of studies that evaluated the same natural experiment. The RD review protocol specified that more detailed analysis of subsets of studies would be conducted if multiple studies were identified that evaluated the same intervention or that investigated sufficiently similar policy questions. The rationale for the sample selection involved two additional considerations. First, it would be easier for a reviewer unfamiliar with RD to appraise multiple studies of the same intervention as each study was likely to have a numerous elements of design and reporting in common. Second, as systematic reviews commonly evaluate the evidence on a single intervention (or group of similar interventions), this sample was more likely to reflect a real-life implementation of a review tool as compared to a random sample of studies on diverse topics.

The steps of the pilot were as follows. Both reviewers ensured familiarity with RD methodology by reading two methodological review articles (Lee and Lemieux, 2010; O'Keeffe et al., 2014) and the WWC standards document. We held an initial meeting to read through the standards together and clarify our understanding of the criteria, instructions for implementation, and the test methodology. We then independently appraised one study (Carpenter and Dobkin 2011) and met to compare our answers. The purpose of this initial appraisal was to clarify any further issues of understanding or interpretation of the tool and to ensure a consistent approach to its application. We then independently appraised the remaining 16 MLDA and met a final time to compare our assessments and discuss our experiences of using the standards. I recorded our individual assessments, reasons for any initial disagreements, and our consensus on final assessments.

5.4.3 Adaptation of WWC Standards and Development of RD-10 Checklist

Based on the findings of the pilot, I decided to adapt the WWC standards into a checklist. The rationale for doing so involved the following considerations: usability; consistency with the design approach of other critical appraisal tools in common use in public health systematic reviews; potential ability to differentiate between higher and lower quality RD studies; and potential ability to identify specific elements of high, low, or uncertain quality in RD studies.

I worked with two additional reviewers to test the usability of the checklist and to investigate how subjective the interpretation of criteria and of study quality might be. Both reviewers were highly experienced in the systematic review of non-randomised studies; in addition, one reviewer had expertise in natural experiments and prior knowledge of RD designs.

As the intention was to use the adapted checklist to assess the quality of RD studies across a wide range of topics in public health, a purposive sample of 13 RD studies was selected in order to include examples of different types of forcing variables (age, date, clinical measurements, and social measures), different interventions or exposures, and different academic disciplines (economics, education, and health). Following an initial meeting to review and discuss the criteria, the reviewers independently appraised the papers, recorded their assessments, and made notes of any queries or problems. When the appraisals were completed, we met to discuss and compare the results.

Measurement of interrater agreement using kappa statistics was considered and rejected for two reasons. First, the checklist was at too early a stage of development; usability and face validity were felt to be sufficient considerations at this stage. Second, almost all of the interrater disagreements stemmed from difficulty in finding the relevant evidence in the paper (particularly lengthy economics papers, whose reporting structure is unfamiliar in public health), meaning that kappa would reflect similarities in ability to find information in the papers rather than similarities in interpreting checklist criteria or similarities in ‘correctly’ answering questions.

5.5 Results

5.5.1 Tools and quality criteria

The literature search did not identify any critical appraisal tools designed for the assessment of RD studies in health. However, the search did identify one published critical appraisal tool designed to assess RD studies in education.

The What Works Clearinghouse (WWC) standards were, at the time of this review, the only published critical appraisal tool for RD. The standards were developed for use in systematic reviews of educational interventions and is published by the US Department of Education. The complete standard involves the application of three criteria to determine whether the study qualifies as a regression discontinuity design followed by ten further criteria to determine whether the study meets four standards. The user then determines whether the resulting combinations of standards mean that the overall WWC standard of evidence for the effects of educational interventions has been met, met with reservations, or not met.

WWC has several strengths. It was developed by a panel that included recognised experts in regression discontinuity designs from the fields of education, economics, and statistics. It is supported by a comprehensive document that explains how to use the standards, elaborates upon the criteria, and explains why they are important. Most of the criteria are specific to features of the RD design. However, WWC also has some limitations. The standards as published in 2010 were produced as a pilot and there is no evidence of user testing or validation. WWC is designed to evaluate studies of educational interventions with pre- and post-test data for individual participants and thus its applicability to public health and health economic studies, which may use administrative and population-level data and may be cross-sectional or retrospective, is unknown. Given the comparative strengths of the tool and the unknown applicability to public health, I decided to make WWC the subject of the pilot.

5.5.2 Pilot of WWC Standards

After the reviewers independently completed the initial appraisal of one MLDA study (Carpenter and Dobkin, 2011), issues with more than half (7/13) of the WWC standards criteria required discussion prior to appraising the full sample of papers. These seven criteria included one of the three eligibility questions plus criteria from all four standards. Table 5.1 describes the criteria, issues encountered, and decisions made to address the issues.

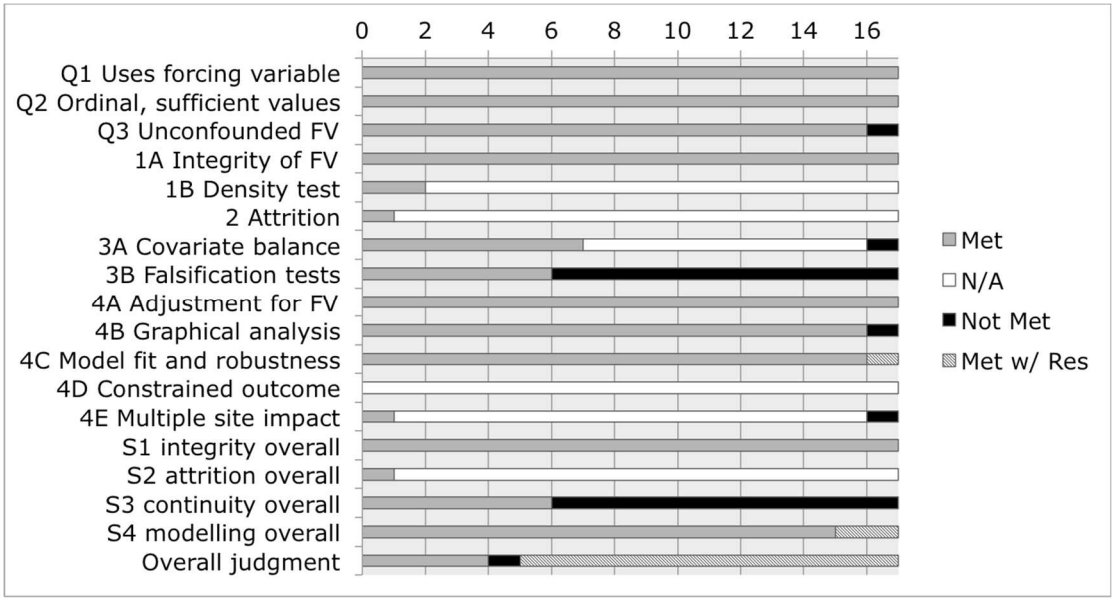
Table 5.1. Issues identified in the pilot of the What Works Clearinghouse Standards for RD for appraisal of studies evaluating the health effects of minimum legal drinking age (MLDA) legislation

Criterion	Issue	Decision
Third qualifying question: Cutoff value must not be used to assign participants to other interventions	The criterion is of critical importance because studies are disqualified if not met. The study did not address this criterion. The reviewers were not aware of other interventions that use age 21 as a cut-off, but also had limited knowledge of American age restrictions or how these might vary between states.	Consider the criterion to be met in the absence of any knowledge or reported information to the contrary. Furthermore, appraise all studies in the sample using all criteria, even if qualifying criteria or standards are not met, in keeping with the purpose of the review to describe study quality rather than to identify a reliable evidence base for an intervention.
Standard 1, Criterion A: "an adequate description of the scoring and treatment assignment process…[which] must show that manipulation was unlikely because scorers had little opportunity or little incentive to change 'true' scores"	The criterion contains seven different information components. The study did not address manipulation of the forcing variable, probably because it used US vital statistics from 1975-1993.	Assume manipulation of age is unlikely and allow the criterion to be met even if manipulation of the forcing variable is not addressed.
Standard 1, Criterion B: Statistical tests or graphical analysis should establish smoothness of the forcing variable around the cutoff	There seemed to be no logical reason for the authors to do this as there would be no opportunity for data subjects or administrators to manipulate birthdates or age.	Modify the standards to allow 'not applicable' as a response. Agreement that the criterion was not met (in the absence of evidence), but that it was neither reasonable nor useful to expect authors to do this for application of the RD design in this situation.

Criterion	Issue	Decision
Standard 2: attrition must be reported to the same standard as an RCT	The study provided no information about attrition. The study was retrospective and reported estimated mortality rates based on death certificates from 1975-1993. Studies that fail to meet this standard also fail the overall WWC standard of evidence.	Modify the standards to allow 'not applicable' as a response.
Standard 3, Criterion A: "Baseline (or pre-baseline) equivalence on key covariates (as identified in the review protocol) should be demonstrated at the cutoff"	The study design and data sources did not allow for investigation of covariates, and the review protocol did not specify any as it was not a review of effectiveness.	Modify the standards to allow 'not applicable' as a response.
Standard 4, Criterion D: empirical support must be provided in the case of any constraints on the values of the forcing variable	The standards document does not say how to judge this criterion if there are no such constraints, yet the standard is not met if the criterion is not satisfied.	Modify the standards to allow 'not applicable' as a response.
Standard 4, Criterion E: specifies reporting of impacts across multiple sites	The standards document does not say how to judge this criterion if it is not a multi-centre study, yet the standard cannot be met if the criterion is not satisfied.	Modify the standards to allow 'not applicable' as a response.

Having agreed these modifications to the standards, the two reviewers independently appraised the remaining 16 MLDA studies. Agreement before discussion was very high, with only seven instances of disagreement. The initial disagreements stemmed from differing prior knowledge of the study context affecting judgment about the unconfoundedness of the forcing variable (n=3), unclear description of the data source in the study (n=1), uncertainty as to whether referencing another publication was acceptable evidence towards the standard (n=1), failing to spot relevant information in the paper (n=1), and erroneous application of the WWC guidance (n=1). Agreement was 100% after discussion. Figure 5.1 presents the appraisal of the MLDA studies after discussion.

Figure 5.1. Results of appraisal of 17 minimum legal drinking age (MLDA) studies using the What Works Clearinghouse Standards for RD



The horizontal axis shows the number of studies falling into each category of judgment (met, not met, met with reservations, or not applicable). The vertical axis shows the qualifying questions (Q), criteria, and standards from the WWC document for which judgments were made.

The pilot demonstrated that the tool could be used on studies in disciplines other than education research, but with modifications required to over half the criteria. Furthermore, the standards were time consuming to use and therefore not feasible to be applied in full to the large number of studies included in this systematic review. The formation of judgments as to whether a study met each standard statement (a combination of met/not met rules for several criteria, some of which had multiple components) and the overall evidence standard (a combination of met/not met rules for the standard statements) was time-consuming and added complexity to the appraisal process without adding value, particularly as the review protocol specified that studies would not be excluded based on quality. More importantly, however, 16/17 studies failed to meet the overall quality standard because they did not report study attrition in the same manner as a randomised trial; all of the MLDA studies were retrospective. I concluded that many studies using population data would fail the standard, which then would not be useful to distinguish differing degrees of quality among studies. However, the individual WWC criteria were easy to apply, most were useful in identifying strengths and limitations of the studies, and interrater variability was low after the initial appraisal and discussion.

5.5.3 Development of RD-10 Checklist

In order to retain these benefits of WWC while improving both useability and applicability to studies in public health and policy, I decided to use the content of the standards as a basis for developing a critical appraisal checklist for RD studies. The checklist approach to critical appraisal is familiar in health sciences and an adaptation would have the following benefits:

- Retain elements of WWC that are applicable to RD studies in health and useful in distinguishing high and low quality studies
- Modify or discard elements of WWC that were of limited applicability in health, caused difficulties of interpretation, or were excessively time-consuming
- Remove complex decision rules, taking a more descriptive than evaluative approach
- Ensure each criterion assesses a single aspect of the study.

The criteria for the adapted checklist (henceforth "RD-10") are shown in Box 5.2.

Box 5.2. RD-10 checklist for critical appraisal of regression discontinuity (RD) studies.

Answer 'yes' or 'no' to indicate whether the criterion accurately describes the study as it has been reported, taking into account any online supplements, appendices, and published protocols. 'Inadequate information' is additionally permitted as a response to criteria 2 and 3.

1. A forcing variable with a threshold or cut-off value is used for treatment assignment
2. The forcing variable is ordinal with at least four unique values on either side of the cut-off
3. An argument is provided regarding the unconfoundedness of the forcing variable
4. A description is provided of the scoring and treatment assignment process that makes the case for the integrity of the forcing variable
5. Smoothness of the density of the forcing variable is established through graphical presentation or a McCrary density test
6. Attrition is described such that it is possible to determine the numbers of participants or observations in the original sample, lost during key stages of analysis, and included in the final analysis
7. The baseline values for key covariates are presented for treatment and control groups
8. Falsification tests of the discontinuity at the cut-off are conducted, either by testing for discontinuities at values of the forcing variable other than the cut-off, or by testing for discontinuities at the cut-off in outcomes that should not be affected by the treatment

9. Robustness checks of the model specification are conducted, such as different functional form specifications or different bandwidths of the forcing variable
10. The statistical model controls for the forcing variable

Table 5.2 shows how the RD-10 criteria map to WWC.

Table 5.2 Comparison of RD-10 to WWC Standards

RD-10	Relationship to WWC standards
1. A forcing variable with a threshold or cut-off value is used for treatment assignment	Based on first qualifying criterion (simplified wording)
2. The forcing variable is ordinal with at least four unique values on either side of the cut-off	Based on second qualifying criterion (more precise wording)
3. An argument is provided regarding the unconfoundedness of the forcing variable	Based on third qualifying criterion, "There must be no factor confounded with the forcing variable"
4. A description is provided of the scoring and treatment assignment process that makes the case for the integrity of the forcing variable	Based on standard 1 criterion A, "The institutional integrity of the forcing variable should be established by an adequate description of the scoring and treatment assignment process"
5. Smoothness of the density of the forcing variable is established through graphical presentation or a McCrary density test	Based on standard 1 criterion B, "The statistical integrity of the forcing variable should be demonstrated by using statistical tests found in the literature or a graphical analysis to establish the smoothness of the density of the forcing variable right around the cutoff"
6. Attrition is described such that it is possible to determine the numbers of participants or observations in the original sample, lost during key stages of analysis, and included in the final analysis	Based on standard 2, which requires that an RD study meet the same standards for reporting of attrition as in a randomised controlled trial
7. The baseline values for key covariates are presented for treatment and control groups	Based on standard 3 criterion A (simplified wording and removal of requirement to demonstrate "equivalence")
8. Falsification tests of the discontinuity at the cut-off are conducted, either by testing for discontinuities at values of the forcing variable other than the cut-off, or by testing for discontinuities at the cut-off in outcomes that should not be affected by the treatment	Based on standard 3 criterion B, which requires either graphical or statistical evidence of no unexplainable discontinuities at values other than the cut-off
9. Robustness checks of the model specification are conducted, such as different functional form specifications or different bandwidths of the forcing variable	Based on standard 4 (Functional Form and Bandwidth), which specifies these two issues as "the most critical aspects of the statistical modelling" but presents five criteria in total
10. The statistical model controls for the forcing variable	Based on standard 4 criterion A (simplified wording and reduction to a single criterion)

Next, the checklist was tested by having two reviewers independently appraise the purposive sample of 13 studies and informally discuss their results. In appraising 13 papers against 10 criteria, 130 judgments could be compared. In

20 of the judgments one of the reviewers gave a response that expressed uncertainty rather than choosing from the permitted responses, and in 11 one of the reviewers gave no response, leaving 99 judgments (76.2%) that could be directly compared. Of these the pairs of reviewers initially agreed on 68/99 (68.7%) of assessments against individual criteria and disagreed on 31/99 (31.3%). These results are displayed graphically in figure 5.2.

Figure 5.2. Agreement between independent reviewers on appraisal of 13 studies using RD-10

ID	RD1	RD2	RD3	RD4	RD5	RD6	RD7	RD8	RD9	RD10
Albouy 2009	Y				N	N				
Bor 2014	Y		I		Y			N		
Chen 2013	Y				N					Y
Decker 2005	Y	Y		Y			Y			
Garrouste 2011	Y	Y	Y	Y	Y		Y		Y	
Johnston 2009		Y	Y	Y			N	Y	Y	N
Jensen 2015	Y			Y					Y	
Lindeboom 2009	Y	Y		Y	Y		N			
Mezuk 2009		Y						N	N	N
Olsho 2015	Y		Y				Y	N	N	Y
Powdthavee 2010	Y	Y		Y	N	N	N	N		Y
Urquieta 2009	Y	Y			N			N		
Ertan Yoruk 2012	Y	Y	Y	Y	N	N	Y	N	Y	Y

Legend:

	Initial independent agreement (Y=yes, N=no)
	Disagreement due to one reviewer expressing uncertainty or no response
	Initial independent disagreement

The first question, on whether treatment assignment was based on a forcing variable, had the highest initial agreement (11/13, 84.6%), followed by the questions on number of unique values and density testing (8/13, 61.5%). Attrition had the lowest number of judgments showing initial agreement (3/13, 23.1%); for seven of these judgments, one of the reviewers did not record any response. There was low initial agreement on the questions relating to the unconfoundedness and integrity of the forcing variable (5/13 and 7/13 respectively) and on questions relating to statistical reporting.

Agreement after discussion was 100%, but discussions were lengthy compared to the WWC pilot. The most common reason for disagreement was that one reviewer did not identify the evidence to support a 'yes' answer when it did exist (sometimes only in a footnote or appendix). The other source of disagreement was how to interpret the relevant evidence in the papers.

Reviewers had questions about how to apply the criteria to different types of forcing variables and different types of RD studies (for example, exploratory or epidemiological studies versus evaluations of the effects of an intervention, as well as studies in which RD is not the main analysis and is reported only briefly). The reviewers also noted that some criteria involve more subjective decisions than others, particularly the assessment of unconfoundedness, the integrity of the forcing variable, and the reporting of attrition.

The conclusion from this test of RD-10 was that it was possible to apply the tool to a range of studies from different disciplines involving different RD designs and forcing variables. The main barriers to implementation were the amount of time required to identify the relevant evidence from the papers being appraised and the amount of time required for discussion. Because of these considerations, it was not feasible to have two reviewers independently appraise all 181 studies in the systematic review of RD reported in chapter 3. Full appraisal results for all studies are available on request as an Excel spreadsheet.

5.6 Discussion

5.6.1 Implications of the findings

This chapter makes a preliminary contribution to the methodological developments required for the critical appraisal of RD studies in health. The findings demonstrate that it is possible to apply a quality assessment tool (WWC) developed for prospective evaluation studies in education within a systematic review of a public health policy topic. However, given that the studies were population-based and retrospective, the quality assessment was largely negative, the standard was overly sensitive to the assessment of attrition, and the tool was not useful in discriminating between higher and lower quality studies. The attempt to adapt the WWC tool into a checklist involving the assessment of individual criteria rather than overall standards produced useful information about the quality of RD studies in health, but the checklist showed limitations in terms of interrater agreement and usability for reviewers less familiar with the technical details of RD design and reporting.

This exercise suggests that a need exists for a critical appraisal tool that is specific to the features of RD designs and describes the quality of these studies in adequate detail, while also being accessible to reviewers who may not have specialist knowledge of RD or familiarity with the reporting conventions of different disciplines, particularly econometrics. Such methodological development is necessary if the aim of incorporating the results of natural experiments in public health systematic reviews is to be realised. The same need and similar challenges may be anticipated for the critical appraisal of other types of natural experimental designs, such as instrumental variable studies. An inability to understand and evaluate such studies can only contribute to their continued exclusion from the public health evidence hierarchy, evidence syntheses, and guidelines.

5.6.2 Reflections on methodology

A strength of the work reported in this chapter is the detailed description of the methodology used, including the reasons for decisions taken at each step of testing. The methodology used broadly conforms to the process described by Deeks et al. (2003), namely preliminary conceptual decisions followed by assessment of face validity, 'field trials' or tests of applicability, and revision (p. 36).

Conformity to elements of good practice in critical appraisal adds further value to this work. RD-10 does not involve a score as this practice is specifically discouraged in the Cochrane handbook on the grounds that it has no empirical basis, produces unreliable assessments, and reduces transparency (Higgins and Green 2011, section 8.3.3). RD-10 also can be seen as an improvement on WWC because each criterion involves only a single question with clear wording. WWC criteria descriptions often involve several different statements and it is not clear how the user should make a judgment if some aspects of the description are met but not others. RD-10 meets the five criteria for tool selection described by Viswanathan et al. (2017), assuming that the expert opinion behind WWC and the applicability testing conducted are sufficient evidence of face validity.

A final strength of this work is the testing of the criteria by using a relatively large sample for dual-reviewer appraisal, followed by application to a large

number of studies in a comprehensive review of RD studies in health. Whereas the applicability of ROBINS-I was tested on a sample of five studies (Sterne et al., 2016) and the Cochrane Handbook states that three to six papers might be a suitable sample for checking consistent application of risk of bias criteria in a review (Higgins and Green 2011, section 8.3.4), in the present work the first stage (testing WWC) involved a sample of 17 studies, the second (testing RD-10) a sample of 13 studies, and the final sample to which the tool was applied consisted of 181 studies.

Poor interrater agreement would appear to be a serious limitation of the RD-10 checklist. Some mitigation of the poor interrater agreement in the present work lies in two explanatory factors: the choice of a purposive sample of RD studies and the technical difficulty of the reports. A further consideration regarding this limitation is that low interrater reliability is a characteristic of critical appraisal tools generally, with studies having reported fair to poor interrater agreement for Cochrane RoB (Armijo-Olivo et al., 2014; Hartling, Hamm, et al., 2013), NOS (Hartling, Milne, et al., 2013), and ROBINS-I (Thomson 2018). This characteristic may be less a product of suboptimal tools or tool development and more an effect of poor reporting in the appraised studies as well as a reflection of the nature of critical appraisal itself. Judgments may differ according to the user's experience with the tool, the study design, critical appraisal methods generally, and the topic under review.

A final limitation of the methodology in this chapter is the absence of a formal consensus approach or more formalised methods of usability testing. The work reported here is therefore preliminary in nature. It could form the basis of further, more formalised development and evaluation of a revised version of RD-10.

5.6.3 Future Developments

The experience of testing RD-10 with different reviewers points to several considerations that should be addressed in any further development of a critical appraisal tool for RD or other natural experimental designs. Lack of familiarity, not only with RD designs in the abstract, but with the reporting of RD studies, particularly from disciplines other than health, is a serious barrier to usability,

interrater agreement, and probably the uptake of any RD critical appraisal tool. A reviewer cannot appraise an RCT without understanding concepts such as sequence generation, allocation concealment, and blinding; knowing how to find the relevant information in a paper; and being able to interpret what is reported in the paper in terms of the appraisal criteria. Similarly, reviewers need knowledge of RD features such as treatment allocation by forcing variable and model specification, along with experience of how these are reported. Consideration is needed as to how to present these features in a way that is accessible to non-expert users, what supporting documentation and training may be required, and how to build capacity among systematic reviewers to conduct such appraisals. All of the initial disagreements between reviewers in the test of RD-10 can be related to inability to identify or locate the relevant information in the paper and/or uncertainty as to how the criterion should be interpreted in the context of the study under assessment. Arguably, then, problems with interrater reliability in critical appraisal are also problems of clarity and transparency of reporting in RD studies. If methods of critical appraisal need development, so too do standards of reporting for RD.

It is also striking that the limited investigations of RD study quality to date have on the one hand focused to a large extent on the quality of statistical reporting, while on the other hand largely neglected more generic yet arguably more important domains of risk of bias, such as measurement bias and selective reporting. The quality of statistical reporting can be considered as distinct from study quality as defined in terms of risk of bias; statistical reporting quality is either absent from most critical appraisal tools in common use in health or restricted to very simple and readily identifiable issues such as the presence of confidence intervals. This exclusion of statistical reporting quality has the benefit of making these critical appraisal tools accessible to non-statisticians, usable within a multidisciplinary review team, and possible to implement with minimal training across a wide variety of fields and topics. The focus on statistical reporting may prove to be a factor limiting the incorporation of natural experimental studies in systematic reviews in health.

Greater knowledge of the empirical relationship between RD design elements and biased estimates of effects would be useful for determining which aspects of

statistical reporting would be most important to include in a critical appraisal checklist and which could be eliminated. If this knowledge were also used to improve reporting by highlighting key aspects of study design in plain language, the result could be a simplified critical appraisal process that could be implemented more easily and more widely.

Further research on this area is envisaged to develop RD-10 into a tool that could be used by systematic reviewers who are not experts in RD methodology, thus helping to realise the potential for natural experiments to be used more widely as evidence in public health. The development process could involve the following:

1. Revision of risk of bias criteria based on experience acquired in the systematic review of RD (chapter 3)

2. Consideration of additional, non-design-specific domains of bias such as selective reporting, including a literature review to identify empirical evidence of such biases with respect to RD studies

3. Consensus-based research such as a Delphi activity involving experts in evidence synthesis, with and without specialist knowledge of RD, to investigate the face validity and acceptability of a revised checklist

4. Development and evaluation of user guidance and training materials to ensure that users can appraise studies relatively quickly and consistently, without being experts in RD

5. Usability testing to ensure criteria can be consistently interpreted and applied

6. Measurement of interrater reliability in an appropriately selected sample of RD studies

7. Field testing in systematic review and guideline development projects to evaluate implementation in terms of feasibility and acceptability

8. Extension of this process and application of learning to development of usable, design-specific appraisal tools for other natural experimental methods, such as instrumental variable, difference-in-difference, and synthetic control studies.

An implementation study would additionally demonstrate real-world utility, further identify user needs and difficulties for revision of user guidance and training materials, and create examples of how appraisal results can be presented and incorporated into systematic reviews.

5.7 Chapter summary

This chapter has reported the testing and development of methods for the critical appraisal of RD studies. The What Works Clearinghouse Standards for RD were applied to 17 public health policy studies and showed limitations in the applicability of criteria to retrospective evaluative designs and in suitability for use in a systematic review. Accordingly relevant criteria from the standards were used as the basis for development of a ten-item checklist, RD-10. Compared to WWC, RD-10 produces a more detailed description of quality and is more applicable to the retrospective RD designs frequently seen in public health research. Issues relating to the reporting quality of RD studies, different conventions of reporting across academic disciplines, and difficulty identifying relevant information in study reports contributed to differences between reviewers in applying the tool.

Chapters 3, 4, and 5 have reported findings from a systematic review of studies using one natural experimental design, namely RD. Chapter 6 will demonstrate a different approach to investigating the contribution of natural experiments to public health evidence by reporting an overview of systematic reviews relating to environmental causes of disease. As described in chapter 2, this is a topic to which natural experiments might reasonably be expected to form part of the evidence base.

6 Endocrine disrupting chemicals and breast cancer risk: A meta-review

6.1 Chapter overview

Whereas chapter 5 considered the application of natural experimental methods to the evaluation of an intervention, chapter 6 looks at the other type of health research question to which these methods may be applied, namely the investigation of the effects of environmental exposures. This chapter begins with a brief introduction to endocrine disrupting chemicals (EDCs), explaining why they are a public health concern, how they have become linked to breast cancer, and how natural experiments might contribute to understanding the environmental causes of disease. The chapter then investigates, via a protocol-driven meta-review or overview of systematic reviews, what evidence has been assembled about EDCs and breast cancer risk and what contribution natural experiments have made to that evidence base. The systematic reviews are brought together in a narrative synthesis which describes in tables the reviews' characteristics and conclusions. The quality of the reviews is assessed and described using the AMSTAR-2 appraisal tool. The findings of the meta-review are presented visually through (1) a diagram of overlap across reviews addressing one group of EDCs and (2) a map of evidence. The chapter concludes by discussing why natural experiments make a very limited contribution and what scope might exist to strengthen this evidence base.

6.2 Aims

By conducting and reporting a meta-review, this chapter aims to answer the following research questions:

1. What is the evidence from systematic reviews that endocrine disrupting compounds (EDCs) cause breast cancer in humans?

2. What is the contribution of natural experiments to the evidence base on the causal role of EDCs in breast cancer?

3. How have systematic reviews evaluated and presented evidence from different study designs, including natural experiments, in reaching their conclusions about EDCs?

4. How do systematic reviews vary in their methodology with respect to inclusion criteria, appraisal methods, and synthesis methods, and how do these variations affect the inclusion and presentation of results from natural experiments?

5. What have systematic reviews identified as the limitations and gaps relating to natural experiments within the evidence base on EDCs and breast cancer in humans?

6.3 Background

6.3.1 Breast cancer: epidemiology and the public health response

Breast cancer is the most common cancer in women worldwide, the leading cause of cancer death among women in developing countries, and the second most common cause of cancer death (after lung cancer) among women in high-income countries (International Agency for Research on Cancer, 2012). In Scotland the incidence of breast cancer has increased over the past two decades and the lifetime risk of breast cancer for women is 11.9%, or approximately 1 in 8 women (Scottish Public Health Observatory, 2018).

Breast cancer is a heterogeneous disease divided into subtypes according to hormone receptor status and HER-2 protein receptor status. The majority of breast tumours (approximately 70%) are oestrogen receptor positive (ER+), meaning that their growth appears to be stimulated by the presence of oestrogen (Macmillan Cancer Support, 2013). It is generally agreed that oestrogens and other hormones play an important role in the aetiology of breast cancer (Trichopoulos et al., 2008). Well-established risk factors for breast cancer include increasing age, reproductive history (age at menarche, age at first birth and parity, breastfeeding, age at menopause), family history, height, birthweight, high body mass index, postmenopausal weight gain, postmenopausal hormone therapy, alcohol intake, exposure to ionising radiation,

and mammographic density (Tamimi, Hankinson, and Lagiou, 2018). Other than weight and alcohol, most of these risk factors unfortunately are not modifiable. Public health messages relating to breast cancer prevention include minimising weight gain, avoiding alcohol, breastfeeding if possible, and physical activity (Tamimi et al., 2016). Increases in breast cancer survival have been achieved through screening programmes that lead to earlier detection and treatment and through healthcare improvement, including more effective treatment and better organisation and delivery of care (Scottish Public Health Observatory, 2018), rather than through prevention.

6.3.2 Endocrine disrupting chemicals as suspected causes of breast cancer

The incidence of breast cancer has increased globally over the past century (Trichopoulos et al., 2008) but with nearly a five-fold difference in rates among countries (Tamimi et al., 2016) and higher risk in urban compared to rural areas (Trichopoulos et al., 2008). A four-fold difference in risk between women in North America and Europe compared to women in China and Japan is not explained by adult diet or reproductive factors (Trichopoulos et al., 2008). The unexplained global rise in incidence combined with unexplained geographical variation has led to questions about possible environmental causes of breast cancer (Tamimi, Hankinson, and Lagiou, 2018).

Among the many environmental exposures that could contribute to breast cancer risk, endocrine disrupting chemicals (EDCs) have attracted particular interest due to the role of hormones in breast cancer aetiology (IOM, 2012). The World Health Organization defines an EDC as “an exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations.” Common EDCs include dioxins, polychlorinated biphenyls (PCBs), organochlorine pesticides (OCPs) such as DDT, herbicides, fungicides, the industrial surfactant perfluorooctanoic acid (PFOA), and consumer product chemicals such as bisphenol A, phthalates, nonylphenols, and polybrominated diphenyl ethers (PBDEs, used as flame retardants) (Gore et al., 2015). Although production of many of these chemicals was eventually banned, they were produced in industrial quantities for decades and are now ubiquitous in the environment

(Gore et al., 2015). Others continue to be in common household and industrial use.

Evidence for the endocrine system effects and any increased cancer risk associated with EDCs varies by chemical and is derived from laboratory, animal, and epidemiological studies (Rodgers et al., 2018). Uncertainty as to the applicability of laboratory and animal studies to human health, the potential for bias and inability to rule out confounding in cohort or case-control studies, and the difficulty of exposure assessment have contributed to controversy over EDCs as a cause of cancer (Brody et al., 2007, IOM, 2012). Additionally, the utility and health-protective effects of many of these chemicals in applications such as increasing crop yield, improving fire safety, and preventing insect-borne diseases must be weighed against the potential risks (Sadasivaiah, Tozan, and Breman, 2007, Shaw et al., 2010). These uncertainties have contributed to conflicting public perspectives on EDCs. For example, Breast Cancer UK's position on EDCs is that they should be regulated on the precautionary principle, recognised as preventable risk factors for breast cancer within national cancer plans, and classified as substances for which no safe exposure threshold can be determined (Breast Cancer UK, 2017). This position contrasts with that of Cancer Research UK, which has responded to public concerns about EDCs with advice that "the evidence linking these chemicals to cancer has generally been poor or inconsistent" (Cancer Research UK, 2016).

6.3.3 Role of natural experiments in identifying environmental causes of diseases

To address the question of how to understand and act upon environmental causes of disease, the Academy of Medical Sciences convened a working group chaired by Sir Michael Rutter (Academy of Medical Sciences, 2007). Their report found that, despite clear evidence of the important role environmental factors play in causing disease, specific knowledge of causal pathways was limited (p. 7). Although RCTs provide the strongest evidence of a causal effect, in most situations random allocation to environmental exposures is neither feasible nor ethical, so the report recommended that researchers should use natural experiments "whenever possible" to assess the effects of environmental factors on disease (p. 13). Furthermore, in order to test causal inferences in different

populations and contexts, the report concluded that "The totality of evidence from all sources should be brought together in order to reach sound conclusions" (p. 8).

The relationship between EDC exposure and breast cancer represents an area of current controversy in which natural experimental methods might be applied in order to support stronger causal inference, for example by leveraging situations such as accidental exposures or variation in geographical proximity to a source of exposure. Therefore, it may be expected that systematic reviews that have attempted to represent the evidence on EDCs and breast cancer will have needed to consider the designs of included studies, how the studies address unmeasured confounding, and how to assess and synthesise findings from studies of varying design. Accordingly, this chapter presents a meta-review on EDCs and breast cancer with a focus on the methods used in systematic reviews for identifying, assessing, and synthesising evidence on the health effects of environmental exposures, on the premise that these reviews will potentially include and synthesise results from natural experimental studies.

6.4 Methods

6.4.1 Protocol and deviations

This meta-review was conducted according to a pre-specified protocol published in the PROSPERO registry. The protocol registration number is PROSPERO 2018 CRD42018089344. Appendix 4 contains the protocol as published in the PROSPERO registry. Deviations from the protocol, with justifications, are reported in Table 6.1.

Table 6.1. Deviations from protocol in the systematic review of endocrine disrupting chemicals and breast cancer risk

Statement from protocol	Deviation and justification
IARC monographs will be searched	IARC monographs are not systematic reviews and would not meet inclusion criteria, therefore the website was not searched.
A data extraction form will be designed and piloted on two systematic reviews (one reviewer will pilot the data extraction form and a second will cross-check the extracted data for accuracy).	A second reviewer was not available for data extraction piloting or checking.

Reviews that include a meta-analysis of risk of breast cancer in humans will additionally [in addition to AMSTAR2] be appraised using the MOOSE (Meta-analyses Of Observational Studies in Epidemiology) checklist.	MOOSE is a reporting guideline, not a critical appraisal checklist. Only AMSTAR-2 has been used in this overview for critical appraisal. AMSTAR-2 has specific questions to assess the quality of meta-analyses and has been designed for reviews of both randomised and non-randomised studies.
The primary qualitative outcome of the review is a map of evidence that demonstrates (1) the number and type of natural experimental studies included in the evidence base and (2) the amount of overlap of included studies among the systematic reviews.	It was not possible to produce a map of evidence matching this description for two reasons. First, the included reviews did not specifically identify any studies as natural experiments. Second, overlap across 15 reviews proved too complex to represent in a single diagram. Instead, the map of evidence demonstrates the number and quality of reviews for each subtopic, following Virendrakumar (2017). Overlap is investigated and described, but due to the number of reviews a diagram of overlap was created only for a subset of reviews as an illustration.
Sensitivity analyses will be conducted by review characteristics, review quality, and inclusion of natural experiments.	There were too few reviews on any EDC subgroup and too little differentiation in review quality to make sensitivity analysis meaningful. Also, no reviews specifically included natural experiments. The results section does identify some patterns of reporting by review characteristic (type of synthesis).

Primary outcome

The primary qualitative outcome of the review specified in the protocol is a map of evidence.

The primary quantitative outcome of the review specified in the protocol is the risk of breast cancer in humans, expressed as relative risk (RR), odds ratio (OR), or hazard ratio (HR), associated with a given exposure to an EDC or combination of EDC under a given set of circumstances, with 95% confidence intervals.

6.4.2 Eligibility criteria

The inclusion criteria for this overview were as follows.

Study type

This overview includes systematic reviews, defined as a study that (1) follows a specific, transparently reported method of retrieving and selecting studies in an effort to comprehensively address its research question and (2) presents the

characteristics and results of included papers in some form of synthesis (quantitative, qualitative, or narrative). This definition was derived from the PRISMA statement. Meta-analyses, rapid reviews, and scoping reviews could be included if they met the above definition of systematic review. Primary studies were not included.

Date

This review includes systematic reviews published on or after 1 January 2003 and whose cut-off date for searches is not earlier than 1 January 2002. The year 2002 was chosen because it was the date of publication of the first Global Assessment of the State of the Science of Endocrine Disruptors (International Programme on Chemical Safety, 2002), at which time only very weak evidence was found to exist of a relationship between EDCs and human health.

Language

No language restrictions were imposed at the search stage. The protocol specified a plan for dealing with records in other languages; however, no documents without an English-language abstract were identified, and no abstracts selected for full-text screening were in languages other than English or French, so no translations were required.

PICO

Included reviews had to address the question of the effect of exposure to EDCs (any chemical or combination of chemicals, any dosage, any timeframe) on the risk of breast cancer in humans. Systematic reviews on a broader topic (such as environmental causes of breast cancer, or effect of EDC exposure on the risk of all cancers) were included if the other inclusion criteria were met and separate results on EDCs and breast cancer were presented. The specific PICO of interest was:

Participants/population

Humans exposed to endocrine disrupting chemicals.

Intervention(s), exposure(s)

The exposure of interest is endocrine disrupting chemicals (see section 6.3.2 above for definition). Environmental, household, and occupational exposures were included. Alcohol and benzene were included in the category of organic solvents as occupational exposures, but individual consumption (of alcoholic beverages or benzene as a component of tobacco smoke) as a route of exposure was excluded. Pharmaceuticals (e.g. hormone therapy) were excluded.

Comparator(s)/control

The comparators could be any variation in exposure (including non-exposure), degree, or timing.

Outcome

Risk of breast cancer.

6.4.3 Search strategy

I searched Medline, Embase, the Cochrane Database of Systematic Reviews (CDSR), Biosys Previews, Scopus, and Web of Science for records dated January 2003 – April 2018. Additionally, Google and OpenGrey were searched for grey literature. The search strategies for the bibliographic databases combined keyword and subject index terms for endocrine disruptors and breast cancer with a filter to identify systematic reviews. The full search strategies are reported in Appendix 5.

6.4.4 Study selection

As per protocol, one reviewer (myself) screened at all stages. The protocol specified that any uncertainty over whether inclusion criteria were met would be discussed with a second reviewer. The only uncertainty that arose concerned reviews with very broad scope, for example with outcomes such as ‘any cancer’ or ‘human health’, which led to discussion with a second reviewer (HT).

However, the uncertainty was resolved by referring back to the inclusion criteria as detailed in the protocol. Decisions with reasons for exclusion were recorded in EndNote.

6.4.5 Quality (risk of bias) appraisal

Included systematic reviews were critically appraised using the AMSTAR 2 checklist, which has been developed to allow for the appraisal of systematic reviews containing evidence from both randomised and non-randomised studies, with or without meta-analysis (Shea et al, 2017). Two reviewers appraised each study independently, compared results, and resolved disagreements through discussion. Each appraiser recorded their assessments on an individual Excel spreadsheet.

6.4.6 Data extraction

A data extraction spreadsheet was designed as described in the protocol. The data extracted from each included review were:

Review characteristics: the citation, year of publication, objectives, search cut-off date, databases searched, inclusion criteria, quality appraisal method, method(s) of synthesis

Details of included studies: number of studies and population numbers included in the review, references of included studies (human populations only), number and date range of other included studies (animal and in vitro), designs of included studies in humans

Details of review findings: EDCs covered, characteristics of EDC exposure covered (doses, timeframes, modifying factors), results of meta-analysis of risk of cancer in humans, numeric estimates of risk from included natural experiments in human populations, results of narrative synthesis, overall assessment of risk of bias and/or certainty of evidence, limitations or gaps noted in the evidence base.

6.4.7 Synthesis

The characteristics of the included reviews were described in tables and grouped by the exposures of interest that they addressed. Information about the contribution of natural experiments and about limitations and gaps in the evidence base were summarised narratively and presented in an 'overview of synthesis' table.

The primary quantitative outcome (risk of breast cancer) was presented in a table, with estimates of risk associated with different exposures presented separately. The primary qualitative outcome (map of evidence) was tabulated in a manner adapted from Sightsavers evidence gap maps (Virendrakumar et al., 2017). These maps classify evidence as strong, weak, or inconclusive. Definitions of strong versus weak evidence vary widely and are arguably less clear-cut when the body of evidence does not include randomised studies. For the adaptation of the evidence map used in this review, I classified the strength of evidence based on characteristics that differed across the reviews and are commonly used as quality criteria, namely whether included studies were prospective or not, whether exposures were measured reliably, and whether the reviews assessed study quality. A judgment of inconclusive was determined when the review authors themselves came to this conclusion, or when insufficient information about study design and quality was provided to support a judgment of strong or weak. This classification is pragmatic but not overly dependent on subjective opinion, and serves as a thumbnail sketch to add some relevant detail to the evidence map.

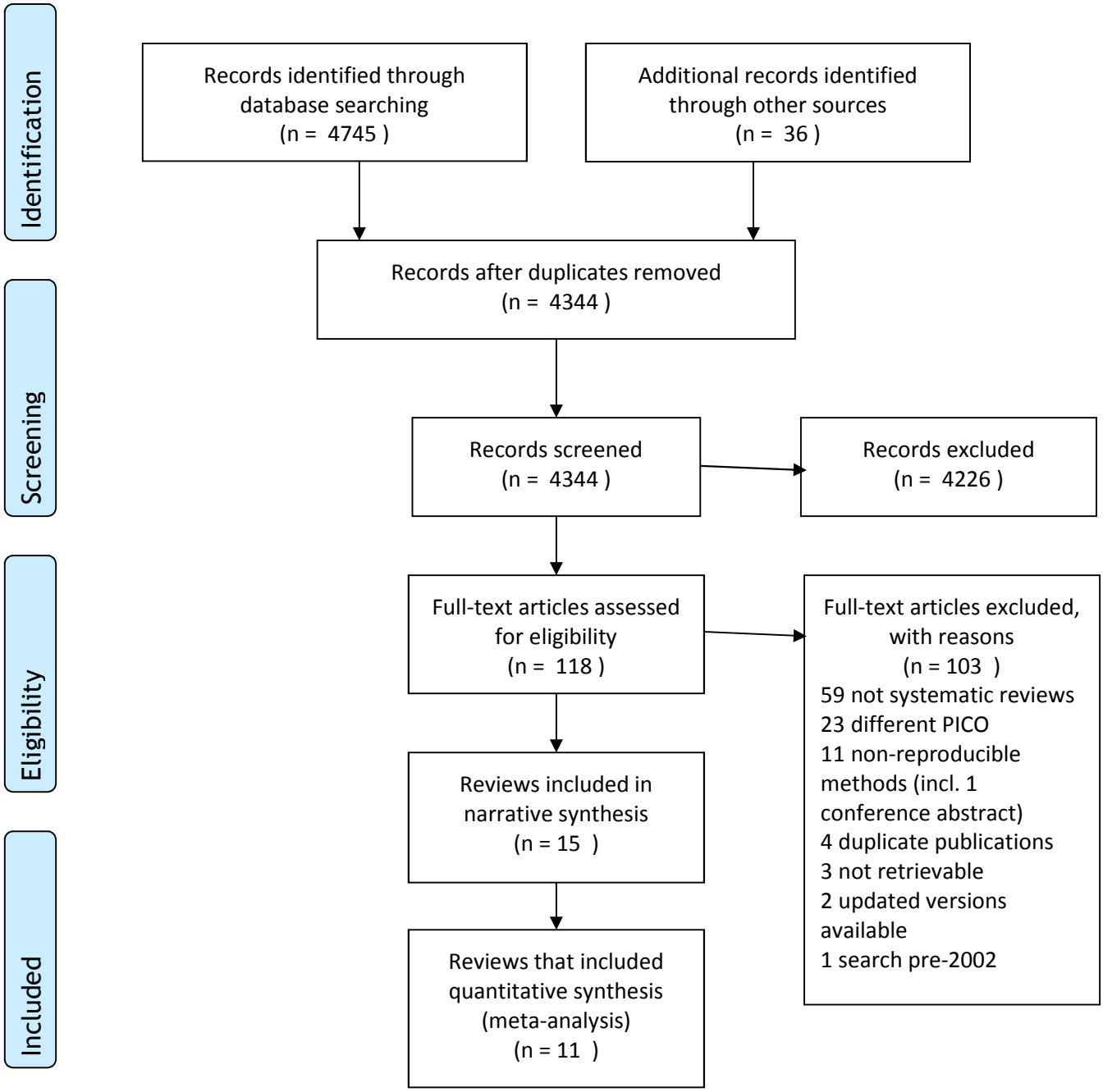
Overlap of primary studies among reviews (Lunny et al., 2017) was determined by cross-tabulating the reviews against a list of references that appeared at least once in each review as part of the evidence base for a similar question. The overlap was then presented graphically as a Venn diagram in a manner derived from McKenzie and Brennan (2017) and summarised in narrative form.

6.5 Results

6.5.1 Literature search results

The search retrieved 4745 records from citation databases and 36 additional records from internet searches. After I removed duplicates and records whose titles and abstracts did not meet the inclusion criteria, 118 full-text articles were retrieved and assessed. Of the full-text articles, 15 systematic reviews met the inclusion criteria; eleven of these reviews contained meta-analyses. Figure 6.1 shows the PRISMA flow diagram for study selection.

Figure 6.1. PRISMA flow diagram for meta-review



6.5.2 Excluded studies

Of the 118 full-text articles that were assessed against the inclusion criteria, 103 were excluded. Of the excluded papers, 59 did not meet even a relaxed definition of a systematic review, 23 did not meet the PICO inclusion criteria, 11 were excluded based on a judgment that the review methods were neither comprehensive nor transparently reported and thus did not meet the definition of a systematic review set out in the protocol, and 10 papers were excluded for other reasons (duplicate publications, not retrievable, updated versions available, date of search outside of included range).

6.5.3 Included systematic reviews

The 15 included reviews were published between 2005 and 2018. Four of the reviews were broad in scope, encompassing a range of EDCs as well as other environmental exposures (Brody 2007, Gray 2017, Mouly 2016, Rodgers 2018). The eleven reviews that contained meta-analyses were more narrowly focused on a group of related chemicals or products. Of these meta-analyses, there were three on pesticides or DDT (Ingber 2013, Khanjani 2007,Park 2014), three on PCBs (Leng 2016, Zani 2013, Zhang 2015), and five on consumer products, including two on hair dyes (Gera 2018, Takkouche 2005), two on deodorants (Allam 2016, Hardefeldt 2013), and one on phthalates (Fu 2017). The four broader reviews also addressed all of these topics, with the exception of Brody 2007, which did not address consumer products. Table 6.2 summarises the characteristics of included reviews.

Table 6.2 Characteristics of included systematic reviews

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
ALLAM, M. F. 2016. Breast Cancer and Deodorants/ Antiperspirants: a Systematic Review. Central European journal of public health, 24, 245-247.	Association between deodorant or antiperspirant use and breast cancer	PubMed, PsycLIT, Current Contents, Best Evidence, cited references, contact with experts Index and keyword terms Actual search strategy not reported Date Database inception to August 2016	Included: n=2 published observational studies on breast cancer risk and deodorant use. English, French, Spanish included. Excluded: Case studies, studies not comparing exposed and unexposed, unpublished studies	None reported	Funding source not stated. COI: none to declare
BRODY, J. G., MOYSICH, K. B., HUMBLET, O., ATTFIELD, K. R., BEEHLER, G. P. & RUDEL, R. A. 2007. Environmental pollutants and breast cancer: epidemiologic studies. Cancer, 109, 2667-711.	Environmental pollutants (either known mammary carcinogens or EDCs) and breast cancer	Pubmed only Yes, table 1 To June 2006 (pesticides 2000-2006, PCBs 1999-2006)	Incl n=152 Peer-reviewed epidemiologic studies in English Excluded: diet, tobacco smoke, certain types of occupational studies, studies with five or fewer exposed women or studies of male breast cancer with <1 observed or expected case	Derived from epidemiology textbook (Aschengrau and Seage, 2003)	Funded by Susan G. Komen for the Cure; no statement of COI
FU, Z., ZHAO, F., CHEN, K., XU, J., LI, P., XIA, D. & WU, Y. 2017. Association	Association between urinary phthalate metabolites and risk of breast cancer	Pubmed, Embase and Cochrane library, keywords only,	cohort studies and case-control studies in English with reported or	Newcastle-Ottawa (scores reported; not discussed)	Funding: Grants from Natural Science Foundation of Zhejiang Province

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
between urinary phthalate metabolites and risk of breast cancer and uterine leiomyoma. Reproductive Toxicology, 74, 134-142.	and/or uterine leiomyoma	to December 20, 2016	calculable RR/OR and CI Excluded: cross-sectional studies, case reports		and Fundamental Research Funds for the Central Universities COI: none to declare
GERA, R., MOKBEL, R., IGOR, I. & MOKBEL, K. 2018. Does the Use of Hair Dyes Increase the Risk of Developing Breast Cancer? A Meta-analysis and Review of the Literature. Anticancer research, 38, 707-716.	Association between personal hair dye use and risk of breast cancer	PubMed, Science Direct, NCBI, keywords 'hair dye' and 'breast cancer' 1980-2017	Epidemiological studies with reported RR/OR and CI, baseline characteristics and selection criteria for cases and controls, and adequate controls with no previous breast cancer diagnosis	None reported	Funding: Breast Cancer Hope Charity No statement of COI
GRAY, J. M., RASANAYAGAM, S., ENGEL, C. & RIZZO, J. 2017. State of the evidence 2017: an update on the connection between breast cancer and the environment. Environmental	"a broad overview of the scientific literature" on exposure to environmental toxicants (including EDCs) and risk of breast cancer	PubMed and Scopus, keywords only, 2007-2017	"epidemiological studies"	None reported	Authors declare no COI Funding "not applicable"

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
Health: A Global Access Science Source, 16, 94.					
HARDEFELDT, P. J., EDIRIMANNE, S. & ESLICK, G. D. 2013. Deodorant use and breast cancer risk. Epidemiology, 24, 172.	Effect of deodorant use on breast cancer development	Medline, Embase, Current Contents Connect, Google Scholar, keywords only, no language restriction, 1950-2012	Published studies with internal control group, controls not diagnosed with breast disease, risk estimate given	None reported	No statement of funding or COI
INGBER, S. Z., BUSER, M. C., POHL, H. R., ABADIN, H. G., EDWARD MURRAY, H. & SCINICARIELLO, F. 2013. DDT/DDE and breast cancer: A meta-analysis. Regulatory Toxicology and Pharmacology, 67, 421-433.	Effect of DDT exposure on risk of breast cancer	PubMed and Web of Science plus cited references, keywords and MeSH index terms, unknown date through June 2012, English language only	Included: studies examining correlation between DDT/DDE exposure and breast cancer risk, with data on both exposure and risk	None reported	Funding: CDC and Oak Ridge Institute for Science and Education COI: None declared
KHANJANI, N., HOVING, J. L., FORBES, A. B. & SIM, M. R. 2007.	Association between cyclodiene pesticide contamination and breast cancer	Medline (PubMed and Ovid) 1966 to July 2006 and Embase (start date	Published cohort, nested case-control, and case-control studies, exposure	None reported	No funding received COI: None declared

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
Systematic review and meta-analysis of cyclodiene insecticides and breast cancer. Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews, 25, 23-52.		not provided, to July 2006) plus reference lists; keyword and index terms	measured in biological samples, reporting RR/OR with CI, or mean differences in exposure; no language restrictions		
LENG, L., LI, J., LUO, X. M., KIM, J. Y., LI, Y. M., GUO, X. M., CHEN, X., YANG, Q. Y., LI, G. & TANG, N. J. 2016. Polychlorinated biphenyls and breast cancer: A congener-specific meta-analysis. Environment International, 88, 133-141.	Women exposed to any of 209 PCB congeners and an outcome of diagnosed breast cancer	PubMed, Science Direct, Proquest, Web of Science, and reference lists Keywords and MeSH terms Inception to 1 January 2015	Published English-language studies which had to have "unequivocal evidence of exposure" to PCB congeners and report RR/OR with CI	Newcastle-Ottawa Scale	National Natural Science Foundation of China and Ministry of Environmental Protection of China COI: none declared.
MOULY, T. A. & TOMS, L. L. 2016. Breast cancer and persistent organic	Humans (male and female), environmental (not occupational)	PubMed, Scopus, Embase, CINAHL, keywords only, 2006-2015	Cohort and case-control studies published in English with direct biological	CASP checklists (results not reported)	Funding: Australian Research Council DECRA fellowship No statement of COI

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
pollutants (excluding DDT): a systematic literature review. Environmental Science & Pollution Research, 23, 22385-22407.	exposure to persistent organic pollutants excluding DDT, O = breast cancer risk		measurements of environmental exposure and clear selection criteria. Occupational exposures, studies without individual measurement of exposure excluded.		
PARK, J. H., CHA, E. S., KO, Y., HWANG, M. S., HONG, J. H. & LEE, W. J. 2014. Exposure to Dichlorodiphenyltrichloroethane and the Risk of Breast Cancer: A Systematic Review and Meta-analysis. Osong Public Health & Research Perspectives, 5, 77-84.	Association between DDT exposure and risk of breast cancer	PubMed and Embase plus reference lists; keywords and index terms; To August 2012	Cohort or case-control studies in English with OR/RR and CI or data for calculation	None reported	Funding: Ministry of Food and Drug Safety, Osong, Korea COI: None declared
RODGERS, K. M., UDESKY, J. O., RUDEL, R. A. & BRODY, J. G. 2018. Environmental chemicals and breast cancer: An updated review of	Human studies of breast cancer and environmental chemicals identified as potential mammary carcinogens	PubMed only Keywords and MeSH terms June 2006-June 2016	Peer-reviewed human studies in English with risk estimate. Excluded: diet, tobacco smoke, shift work, pharmaceuticals, metals, natural	Assessed studies in terms of selection, exposure assessment, control for confounding, misclassification of exposure or outcome	Funded by Silent Spring Institute and Avon Foundation for Women Authors are employed by Silent Spring Institute

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
epidemiological literature informed by biological mechanisms. Environmental Research, 160, 152-182.			disasters, certain types of occupational studies, studies with five or fewer exposed women or studies of male breast cancer with <1 observed or expected case		
TAKKOUCHE, B., ETMINAN, M. & MONTES-MARTINEZ, A. 2005. Personal use of hair dyes and risk of cancer: a meta-analysis. JAMA, 293, 2516-25.	Personal use of hair dyes and risk of cancer	Medline, Embase, LILACS, ISI Proceedings, article reference lists Subject headings and keywords, no language restrictions, Inception to 2004 (Medline to January 2005)	Published cohort or case-control studies with RR and CI or data for calculation Excluded: occupational exposure to hair dyes	Assessed using a 10-point scale adapted from an unrelated meta-analysis; results not reported	COI: none reported Funding: Canadian Institutes of Health Research postdoctoral fellowship
ZANI, C., TONINELLI, G., FILISETTI, B. & DONATO, F. 2013. Polychlorinated biphenyls and cancer: an epidemiological assessment. Journal of Environmental Sciences and Health, Part C, 31, 99-144.	PCB exposure and risk of any cancer	PubMed only Keywords "1970s" to end of 2012	Published, peer-reviewed cohort and case-control studies of known direct exposure, occupational exposure, or with individual measures of exposure, with OR/RR/SMR or sufficient data for calculation, and English language only. Excluded: Ecological and	None reported	No statement of funding source or COI

Study	Scope/PICO	Searches	Inclusion/ exclusion criteria	RoB assessment method	Review funding source and COI
			cross-sectional studies, cohort or case-control with indirect measure of exposure		
ZHANG, J., HUANG, Y., WANG, X., LIN, K. & WU, K. 2015. Environmental Polychlorinated Biphenyl Exposure and Breast Cancer Risk: A Meta-Analysis of Observational Studies. PLoS ONE, 10, e0142513.	Association between PCB exposure and breast cancer risk	PubMed, EMBASE, CBM and CNKI [Chinese language] databases plus reference lists. Keywords and MeSH terms. Yes, detailed strategies provided ? to November 2014	Cohort or case-control studies, biological samples, RR/OR and CI reported. English and Chinese languages only. Excluded: no biomarker data; fewer than 50 breast cancer cases	Newcastle-Ottawa Scale	Funded by National Natural Science Foundation of China COI: none declared

The included reviews aimed to be comprehensive within their defined scope. All limited their inclusion criteria to studies in humans and the meta-analyses were additionally restricted to studies that quantified risk. Seven of the fifteen reviews restricted the included study types to cohort and case-control studies and three others included only ‘epidemiological’ studies. The searches generally were comprehensive, although three reviews (Brody, Rodgers, and Zani) searched only one database (PubMed) and five reviews (Fu, Gera, Gray, Hardefeldt, Mouly) reported using only keywords in the search strategy (as opposed to a combination of keyword and subject heading or index terms). Three reviews (Allam, Hardefeldt, and Zani) did not report a funding source and five (Brody, Gera, Hardefeldt, Mouly, and Zani) did not provide a declaration of conflicts of interest (COI).

6.5.4 Quality of included systematic reviews

Assessed against the AMSTAR-2 tool, the quality of the included systematic reviews was low (Leng, Takkouche) or critically low (the 13 other reviews). Rating overall confidence according to the number of critical flaws and other weaknesses, none of the reviews could be rated as high or moderate quality according to the AMSTAR-2 guidance, as both of these ratings require no critical flaws (a full ‘yes’ to each of 5 criteria for narrative reviews and 7 criteria for meta-analyses). No reviews mentioned a protocol, which is a critical flaw and immediately drops the overall confidence rating to ‘low’. Any additional critical flaw drops the overall confidence rating to ‘critically low’. Of the 13 reviews in this category, 10 reviews did not provide a list of excluded studies with justifications, 8 did not assess risk of bias, and 3 did not meet the quality criteria for literature searching. The overall AMSTAR-2 appraisal results for each study are presented in Table 6.3. Detailed appraisal results are presented in Appendix 6.

Table 6.3. Overall quality assessment of included systematic reviews on endocrine-disrupting chemicals and risk of breast cancer

Study	Overall confidence	Domains in which study shows critical flaws
Allam 2016	Critically low	Protocol, list of excluded studies, risk of bias assessment, publication bias assessment
Brody 2007	Critically low	Protocol, comprehensive search, list of excluded studies
Fu 2017	Critically low	Protocol, list of excluded studies
Gera 2018	Critically low	Protocol, risk of bias assessment
Gray 2017	Critically low	Protocol, list of excluded studies, risk of bias assessment
Hardefeldt 2013	Critically low	Protocol, comprehensive search, list of excluded studies, risk of bias assessment, appropriate meta-analysis, publication bias assessment
Ingber 2013	Critically low	Protocol, risk of bias assessment
Khanjani 2007	Critically low	Protocol, risk of bias assessment
Leng 2016	Low	Protocol
Mouly 2016	Critically low	Protocol, list of excluded studies
Park 2014	Critically low	Protocol, list of excluded studies, risk of bias assessment
Rodgers 2018	Critically low	Protocol, comprehensive search, list of excluded studies
Takkouche 2005	Low	Protocol
Zani 2013	Critically low	Protocol, comprehensive search, list of excluded studies, risk of bias assessment, publication bias assessment
Zhang 2015	Critically low	Protocol, list of excluded studies

6.5.5 Overview of synthesis

Data were extracted from included reviews as described in the protocol. Studies were grouped by synthesis type (narrative or meta-analysis) and then by the EDCs investigated. The narrative reviews took a broad approach to EDCs compared to the meta-analyses, each of which had a narrower focus, included more studies, and reported more specific conclusions regarding risks associated with the chemical(s) under review. Table 6.4 provides an overview of the

synthesis, including a summary of the evidence base presented in each review and extracts that summarise the review's conclusions.

Table 6.4. Overview of synthesis of systematic reviews

Type of synthesis	Endocrine disruptor	Included reviews	Evidence base, including natural experiments	Risk of breast cancer associated with exposure
Narrative reviews	Environmental chemicals (any)	Brody 2007	152 epidemiological studies assessing a wide range of chemicals (not exclusively EDCs), including 5 studies in which exposure to OCPs was defined by proximity to treated crops, 1 industrial accident (Seveso), 1 industrial contamination (Chapaevsk), 3 cohorts of herbicide workers exposed to TCDD contamination, and 1 study of perchloroethylene-contaminated drinking water	"The strength of evidence…supports an association between PCBs…and breast cancer risk in the 10% to 15% of women who carry certain genetic variants." "Lack of evidence for an association between OCPs and breast cancer may be due to a true lack of association or to shared methodological weakness across a large number of studies." "The evidence regarding dioxin and breast cancer is thus far inconclusive." (p. 2706)
		Gray 2017	"Hundreds" of studies assessing a wide range of environmental exposures (not exclusively EDCs; studies not tabulated). Description of included studies was inconsistent but included the Seveso industrial accident cohort and two studies of a cohort of German factory workers exposed to high levels of TCDD (dioxin).	"The growing literature on developmental exposures to EDCs and later development of breast cancer is especially strong." "the breadth and strength of the evidence cited in this review, when taken as a whole, reinforce the conclusion that exposures to a wide variety of toxicants – many of which are found in common, everyday products and byproducts – can lead to increased risk for development of breast cancer." (p. 42)
		Rodgers 2018	151 epidemiological studies (published 2006-2016) and 7 meta-analyses assessing a wide range of chemicals (update of Brody 2007), including continuing follow-up of Seveso and German factory cohorts. Several studies assessed exposure through geographic location (e.g. residence in a contaminated area, proximity to a factory), which may constitute natural experiments	"New epidemiological studies add to evidence that EDCs and chemicals that are mammary carcinogens in animal models influence the risk of breast cancer." (p. 175) "A precautionary approach is especially important because study methods are limited, short of a 50-year study, to evaluate the life-long risks to humans from these chemicals" (p. 176)
	POPs excluding DDT	Mouly 2016	14 case-control studies on PCBs, OCPs, PBDE, or perfluorinated compounds and 1 cohort on dioxin (Seveso)	"Epidemiological studies published in the last 10 years could neither prove nor rule out the association between breast cancer risk and

Type of synthesis	Endocrine disruptor	Included reviews	Evidence base, including natural experiments	Risk of breast cancer associated with exposure
				environmental exposure to POPs (other than DDT)." (p. 22403)
Meta-analyses	Cyclodiene pesticides	Khanjani 2007	21 case-control studies investigating 10 different chemicals. No natural experiments (based on description of study recruitment methods)	"Our meta-analysis did not show a significant association between any cyclodiene chemical and breast cancer except heptachlor, but that was based on only two studies" [ratio of geometric means 5.32 (95% CI: 3.79 to 7.48); total 305 cases and 340 controls]
	DDT	Ingber 2013	37 case-control studies No description of study context or exposure mechanisms	OR 1.04 (95% CI:0.94 to 1.15) I^2 31.72, p=0.02, possibly due to inconsistent adjustment for confounding across studies "The results of our meta-analysis do not support an association between DDT and DDE exposure and the risk of breast cancer."
		Park 2014	37 case-control studies (11 nested, 15 hospital-based, 11 population-based). No natural experiments (based on description of study design and recruitment)	OR 1.03 (95% CI:0.95 to 1.12) I^2 40.9, p=0.006, possibly due to confounding or effect modifiers "our meta-analysis found no evidence that there is an association between exposure to DDE and the risk of breast cancer"
	PCBs	Leng 2016	16 case-control studies (5 nested). No natural experiments (based on description of study recruitment methods)	The congener-specific meta-analysis found increased risk of breast cancer associated with three of the nine PCB congeners evaluated in two or more studies (eight congeners were only evaluated in single studies). Increased risk was associated with PCB 99 (OR: 1.36; 95% CI: 1.02 to 1.80), PCB 183 (OR: 1.56; 95% CI: 1.25 to 1.95) and PCB 187 (OR: 1.18; 95% CI: 1.01 to 1.39).
		Zani 2013	14 case-control and 9 cohort studies (12 and 6 in pooled analysis). One possible natural experiment (one study conducted in an area contaminated by PCB manufacturing)	OR 1.15 (95% CI:0.92 to 1.43) I^2 70.6%, p=0.000 "The summary ORs …do not suggest a significant association of PCBs with breast cancer, although a

Type of synthesis	Endocrine disruptor	Included reviews	Evidence base, including natural experiments	Risk of breast cancer associated with exposure
				modest effect cannot be entirely excluded. …Overall, epidemiological research yields no evidence for an association between PCB exposure and breast cancer" (pp. 133-134)
		Zhang 2015	25 case-control studies (9 nested). No natural experiments (based on narrative synthesis)	OR 1.09 (95% CI:0.97 to 1.22) I^2 55.4%, p=0.000 Breast cancer risk is associated with groups II and III PCBs but not group I or total PCB exposure (p. 11)
	Deodorant	Allam 2016	2 case-control studies No natural experiments	OR 0.40 (95% CI:0.35 to 0.46) Heterogeneity not assessed Antiperspirant use "could be a protective factor"…"our systematic review did not reveal any possible association"
		Hardefeldt 2013	2 case-control studies No natural experiments	OR 0.80 (95% CI:0.50 to 1.28) (different in figure, which includes 3 studies) Heterogeneity not assessed "We found no evidence from the combined published studies that deodorant promotes development of breast cancer"
	Hair dye	Gera 2018	8 studies No natural experiments	OR 1.1465 (95% CI:0.9962 to 1.3194) (random effect model, not weighted) I2=73.89 (reasons not explored) "the personal use of hair dyes may be associated with an increased risk of breast cancer. …Our findings do not represent evidence for the presence of a cause-effect relationship."
		Takkouche 2005	12 case-control and 2 cohort studies No natural experiments	OR 1.06 (95% CI:0.95 to 1.18) (random effects) Q test <0.001, Ri 0.68 (moderate to large heterogeneity, disappears if Jordanian study excluded) "we did not find strong evidence of a marked increase in the risk of cancer among personal hair dye users"

Type of synthesis	Endocrine disruptor	Included reviews	Evidence base, including natural experiments	Risk of breast cancer associated with exposure
	Phthalates	Fu 2017	3 case-control and 1 cohort study No natural experiments	OR 0.96 (95% CI:0.80 to 1.14) I2=53.30%, p=0.001 (NS when Mexican study excluded) No significant association overall between urinary phthalate metabolites and risk of breast cancer (subgroup analyses associated risk or protective effect with specific metabolites)

This meta-review presents a substantial body of evidence from 15 systematic reviews to address the question of whether EDCs are associated with increased risk of breast cancer. All of the meta-analyses that considered total exposure to groups of related chemicals or consumer products did not find statistically significant increased risk associated with that group. However, some meta-analyses that examined specific chemicals in subgroup analyses did find statistically significant increased risks associated with specific phthalate metabolites or types of PCBs. Also, two of the narrative reviews concluded that the strength of evidence was generally in favour of increased risk and two found that the evidence was inconclusive. The following section describes these findings in more detail by group of chemical or consumer product.

6.5.6 Evidence from systematic reviews on endocrine disrupting compounds (EDCs) and risk of breast cancer

This section answers review question 1, what is the evidence from systematic reviews that endocrine disrupting compounds (EDCs) cause breast cancer in humans?

6.5.6.1 Pesticides

Three meta-analyses (Ingber, Park, and Khanjani) identified a total of 43 case-control studies on pesticide exposure and breast cancer, with no one meta-analysis including all the studies (see analysis of overlap, section 6.5.7). Although the pooled odds ratios were all slightly above 1, the confidence intervals all included 1 (no statistically significant difference in risk of breast cancer). Heterogeneity was statistically significant, which Ingber et al. and Park et al. both attributed to confounding. All three meta-analyses concluded that the evidence did not support an association between DDT exposure and breast cancer, a view shared by Mouly et al. after reviewing 14 of the DDT case-control studies in a broader narrative review on POPs.

6.5.6.2 PCBs

Three other meta-analyses (Leng, Zani, and Zheng) identified a total of 30 case-control studies on PCB exposure and breast cancer, with no one meta-analysis

including all the studies (see analysis of overlap, section 6.5.7). In these reviews, the analysis of PCBs was undertaken in one or more of three different ways: total PCB exposure, exposure by PCB group (I, II, or III), or exposure by specific PCB congener. The conclusions about PCBs and breast cancer risk differ according to the type of analysis undertaken, with small, statistically significant increased risks associated with some groups and congeners, but not others. Zani and Zheng assessed total PCB exposure. The pooled odds ratios were 1.15 and 1.09 respectively, but the confidence intervals included 1 (no statistically significant effect) and heterogeneity was high.

6.5.6.3 Consumer products

Five meta-analyses investigated exposure to various consumer product chemicals and did not find statistically significant increases in breast cancer risk. Two meta-analyses (Allam and Hardefeldt) pooled odds from the same two case-control studies of deodorant use, yet arrived at different results, with Allam reporting pooled OR of 0.40 (95% CI 0.35 to 0.46) and Hardefeldt reporting 0.80 (95% CI 0.50 to 1.28). Examining original study reports is outwith the scope of this overview, but in the absence of evidence to support the biological plausibility of a protective effect, Hardefeldt is more likely to be correct. In any event, both reviews concluded that there was no evidence of an association between deodorant use and increased risk of breast cancer.

Other consumer product chemicals assessed by the included reviews were hair dyes and phthalates. Takkouche and Gera conducted meta-analyses of hair dye use 13 years apart. Both found slightly increased odds of breast cancer in their pooled analyses but these did not reach statistical significance and showed high heterogeneity. Finally, Fu et al. identified one cohort and three case-control studies that assessed urinary phthalate metabolites and breast cancer risk. Although an increased risk or a protective effect were seen with specific metabolites, for total phthalate exposure there was no statistically significant effect and the point estimate suggested a protective effect.

Overall, the 11 included meta-analyses found little or no evidence of increased breast cancer risk associated with exposure to the EDCs studied except for some PCBs. However, the four narrative reviews were less willing to reject the null

hypothesis. Brody, Gray, and Rodgers all tended in their narrative synthesis to emphasise or cite studies that reported increased risks; Gray and Rodgers both stated in their conclusions that EDCs increase the risk of breast cancer, while Brody and Mouly were equivocal (see table 6.4 for quoted extracts).

In comparing the findings of all 15 included reviews, it is important to note some sources of heterogeneity in this overview. The findings of the narrative reviews by Brody, Gray, and Rodgers are not strictly comparable to the other 12 reviews because, despite the stated inclusion criteria, these authors also integrated findings from selected in vitro and animal studies in what appears to be an ad hoc manner and incorporated these into their narrative of the body of evidence in humans. As the examination of overlap will show, these reviews were not as comprehensive as the meta-analyses of DDT or PCBs, missing a considerable number of relevant studies. Additionally, Brody, Gray, and Mouly were less systematic in their data extraction and in the organisation and presentation of their synthesis compared to Rodgers and to the better-quality meta-analyses. Finally, the type of synthesis method combined with the difference in scope appears to have created a fundamental divide between the included reviews, with none of the meta-analyses concluding that risk was increased except in subgroup analyses of specific congeners or metabolites.

6.5.7 Analysis of overlap

Because of the differences in scope of the various included reviews, an analysis of overlap across all 15 reviews would be neither feasible (because of the included number of reviews and cited studies) nor informative (as there is no reason to expect overlap between reviews that focus on mutually exclusive subgroups of EDCs). Two of the meta-analyses had nearly identical review questions and inclusion criteria (Ingber 2013 and Park 2014); therefore, these two reviews were selected to form the basis of an analysis of overlap relating to included studies on DDT/DDE exposure, including for comparison purposes two of the narrative reviews that also examined this exposure (Gray 2017 and Rodgers 2018). (Venn diagrams are usually limited to two or three sets and become very complex to draw when more than four sets are involved.) The overlap is depicted in figure 6.2.

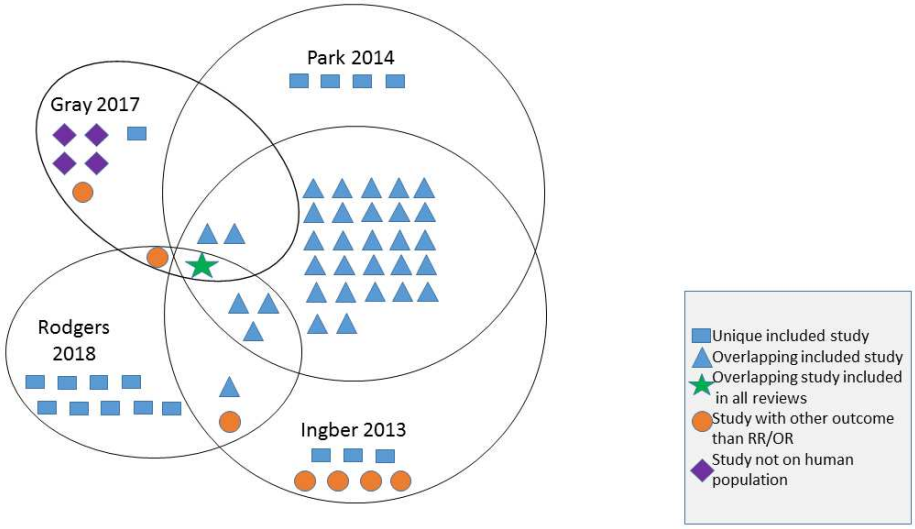


Figure 6.2. Overlap of primary studies included in systematic reviews of DDT/DDE exposure and risk of breast cancer

The visualisation of overlap is informative about the comprehensiveness of each review and about the overall evidence base on DDT/DDE and breast cancer. The Ingber and Park meta-analyses identified the largest proportion of the available evidence; both included 37 studies in their respective meta-analyses, yet overlap in these two similar meta-analyses, published within a few months of one another, was not 100%. In fact, the two meta-analyses have 27 included studies in common. Park additionally identified four unique studies which do not appear in any of the other three reviews and Ingber identified three unique studies. Date of publication also contributes to a lack of overlap, with the two later reviews by Gray and Rodgers contributing a total of ten additional studies to the evidence base (one study unique to Gray and nine unique to Rodgers). In some cases a lack of overlap should be recognised to be caused by differences in (or deviation from) inclusion criteria, with Gray citing four animal or in vitro studies, and all the reviews except Park citing some studies that do not provide a risk difference as an outcome. Overall, in a cited evidence base of 58 studies in human populations and four animal or laboratory studies, only a single case-control study was cited by all four reviews. It is also striking that all four reviews contributed unique and relevant studies to the evidence base.

Overlap was also investigated among reviews that synthesised evidence on phthalates, hair dyes, and PCBs. Fu et al. (2017) included four case-control studies on phthalates, two of which were included in the Rodgers review. Gera et al. (2018) included eight studies in their meta-analysis of hair dyes, of which six were published early enough to potentially be included in Takkouche et al. (2005). Five of those six studies were included, but Takkouche identified an additional 9 case-control studies that Gera et al. appear to have missed or excluded. In the narrative reviews that discussed hair dyes, Gray et al. (2017) only cite one study (also cited by Gera et al.), and Rodgers cites two, one of which is unique to Rodgers. Finally, comparing the three meta-analyses on PCBs (Zhang, Zani, and Leng), Zhang includes 25 studies, which encompass all studies from Zani plus two Chinese-language studies, one of which is a thesis. Leng includes 16 studies, 11 of which are included in Zhang. Of the five studies not included in Zhang, one is listed in Zhang's exclusion table as meeting their exclusion criterion of <50 cases and four are unique to Leng.

Because Leng provided a supplement which lists the excluded studies and reasons for exclusion, it is possible to investigate reasons for the gaps in overlap. Of the 14 studies included in Zhang but not in Leng, nine were excluded by Leng because the analysis was of total PCB exposure and not congener-specific, and two were excluded because the genetic polymorphism analysed was not within scope. The two Chinese-language studies included by Zhang would have been missed due to the English-language restriction in the Leng review. One study that is missing from Leng but present in Zhang (Ward 2000) is indexed in PubMed, is in English, and discusses PCB congeners in the abstract, so presumably represents an error in either the literature search or the study screening process.

6.5.8 Natural experiments in the evidence base on EDCs and breast cancer

This section answers review question 2, what is the contribution of natural experiments to the evidence base on the causal role of EDCs in breast cancer? and review question 3, how have systematic reviews evaluated and presented evidence from different study designs, including natural experiments, in reaching their conclusions about EDCs?

Based on 15 systematic reviews, the contribution of natural experiments to the evidence base on EDCs as a cause of breast cancer is very limited, if not non-existent. None of the reviews specifically identified or described any of the included studies on EDCs as natural experiments. As described in table 6.4, the 11 meta-analyses draw entirely on cohort and case-control designs and offer limited information about the included studies, with the exception of Zani et al. 2013, which provided a detailed narrative synthesis as well as meta-analysis (discussed further below). The only distinction among study designs with reference to causal inference was the distinction made in some, but not all, reviews between nested (within a prospective cohort study) and retrospective case-control designs. Generally the meta-analyses described their results conservatively as providing evidence of association, not causation.

The reviews that used narrative synthesis offered more detailed information about study characteristics and thereby furnished more examples of studies based on exposures that could potentially be analysed as natural experiments. All four narrative reviews described cohort studies based on an industrial accident in Seveso, Italy in 1976, in which a chemical plant explosion exposed people to dioxin (TCDD). Rodgers et al. (2018) note that the Seveso Women's Health Study is "of particular interest" (p. 156) because it involves exposure to one specific substance rather than a mixture and uses an unexposed group for comparison. The most recent findings from this study reported in Rodgers et al. show no statistically significant increase in breast cancer risk in the exposed group, but follow-up of this cohort continues and women exposed earliest in life are only now entering their fifth decade. Other industrial accidents or exposures mentioned in the narrative reviews include dioxin contamination from a chemical plant in Chapaevsk, Russia (Revich 2001, cited in Brody et al. 2007) and perchloroethylene exposure in Cape Cod, USA caused by a fault in drinking water distribution pipes (Aschengrau 2003, cited in Brody et al. 2007).

As described in chapter 2, geographical locations may be used to identify natural experiments, whether by using distance from a relevant location as an instrument or by using geographic boundaries or other features to reliably differentiate exposed and unexposed groups. The narrative synthesis of Zani 2013 described one case-control study that was conducted in Slovakia (Pavuk

2004) in an area contaminated by PCB manufacturing and another in Mexico City where PCB-containing equipment was identified (Lopez-Carillo 2002), which could constitute natural experiments depending on how selection and exposure were assessed (information not provided by the review). The reviews by Brody et al. (2007) and Rodgers et al. (2018) identify a dozen studies in which exposure has been assessed by geographic location, such as residence at a hazardous waste site or inside an industrial park where chemicals are manufactured. However, without reference to the original articles (which is outside the scope of this meta-review), the description of these studies is not detailed enough to determine whether these are natural experiments or ecological studies.

One final observation illustrates how some relevant natural experiments could be missed by reviews focusing on risk of breast cancer as the outcome. One meta-analysis that addresses PCBs and all cancers in humans mentions two situations that could constitute natural experiments. The Yusho incident in Japan (1968) and Yucheng incident in Taiwan (1979) involved mass poisonings caused by rice oil accidentally contaminated by PCBs and polychlorinated dibenzofurans (Zani et al. 2013). Follow up of exposed cohorts found statistically significant increased risks of liver, stomach, lymphatic, and lung cancers, but the review does not report any findings from these cohorts on breast cancer risk. If such cohorts are followed up for all cancers, but no increased risk of breast cancer is identified, this could be an important source of negative findings which may have been missed out from syntheses that focus on reports of breast cancer risk only.

6.5.9 Identification of limitations and gaps within the evidence base

This section answers review question 5, What have systematic reviews identified as the limitations and gaps relating to natural experiments within the evidence base on EDCs and breast cancer in humans?

None of the 15 included reviews comments specifically on limitations or gaps relating to natural experiments, but the reviews do offer some observations about the evidence base and recommendations for future research that are relevant to this question. It is noteworthy that none of the reviews recognised a

need for natural experimental studies; the closest such comments might be Brody's suggestion that further research on OCPs "should be a priority only when researchers have access to novel data that resolves earlier methodological problems" (p. 2706) and Rodgers' recommendation that "more appropriate comparison groups are needed to avoid confounding by differences in baseline risk" (p. 175). By contrast, four reviews stated that large prospective cohort studies were required (Allam, Gera, Gray, and Rodgers) and three called for larger and/or more studies without specifying a design (Fu, Hardefeldt, Leng). More specific gaps identified in the reviews include the need for studies to address interactions between chemicals and interactions with genetic polymorphisms (Khanjani, Leng, Mouly, Rodgers). Two reviews noted a lack of evidence examining risk by breast cancer type or hormone receptor status (Brody, Leng). Finally, a gap was noted in the ability to assess dose-response effects from the evidence base (Leng).

The reviews were more forthcoming on the limitations of the included primary studies. The possibility of confounding was raised in most of the reviews; authors noted inconsistency and limitations across primary studies in adjusting for confounders, including known risk factors for breast cancer, as well as cancer-related weight loss (which can change the concentration of chemicals and metabolites in the body). Further limitations noted in the evidence base related to exposure assessment methods (Brody, Gray, Mouly, Zhang), including insufficient information on age at exposure or differentiation between early- and later-life exposures (Mouly, Park), misclassification due to changes in biological concentrations over time and lag time for cancer development (Zani), the possibility of recall bias in retrospective studies, and varying definitions of exposure categories across studies (Ingber).

6.5.10 Map of evidence

The map of evidence (Figure 6.3) offers a visual summary of the availability and quality of systematic reviews and meta-analyses on the topic area. Note that the map does not represent effect size or direction. Where this information was available from the reviews, it has been reported in Table 6.4.

Figure 6.3 Map of evidence: endocrine disrupting chemicals and risk of breast cancer

Strength of evidence	DDT	OCPs (other than DDT)	PCBs	Dioxin (TCDD)	Bisphenol A	Phthalates	Per- and poly-fluorinated compounds	Organic solvents	Household/ consumer products
Strong			● ○			○			●
Inconclusive	○ ○ ○	○ ○	○ ○	○		○	○	○	
Weak					○				○ ○ ○ ○

Each dot represents a systematic review dedicated to the topic. A larger dot represents a review dedicated to the topic; a smaller dot indicates that a subsection of one or more broader reviews addresses the topic. The quality of the systematic review is represented by the shading of the dot. No reviews were assessed as high quality so reviews have been categorised as either moderate quality (black dot; one critical flaw according to AMSTAR-2 criteria, namely lack of a reported protocol) or low quality (white dot; more than one critical flaw as per AMSTAR-2).

The strength of evidence within the systematic reviews has been categorised as ‘strong’ if (a) it includes evidence based on prospective follow-up (including nested case-control studies) and (b) the review authors assessed the quality of included studies. The strength of evidence is categorised as ‘inconclusive’ if the authors provide insufficient information about the design and quality of included studies, do not reach any clear conclusion about the body of evidence, or specifically state in their review that the evidence is inconclusive. The strength of evidence is categorised as ‘weak’ if (a) the review is based solely on retrospective, cross-sectional, or animal/laboratory evidence or (b) the included studies were assessed by the review authors as poor quality overall.

The map of evidence shows that existing systematic reviews address all of the categories of EDCs covered in this overview (the excluded categories of pharmaceuticals and polyaromatic hydrocarbons were additionally addressed within the broad narrative reviews). However, nearly all of this systematic review evidence is either inconclusive or weak. The strongest evidence comes from two reviews of PCBs, one of phthalates, and one of hair dyes. Despite a substantial body of primary studies on DDT, the strength of evidence is inconclusive because the systematic reviews did not provide sufficient information about the design or quality of included studies. Evidence from systematic reviews is lacking on breast cancer risk associated with bisphenol A, PFASs, and flame retardants such as PBDE used in household products.

6.6 Discussion

This meta-review identified 15 systematic reviews that assemble evidence on EDCs and risk of breast cancer and found that natural experiments contribute little, if anything, to the body of evidence. The quality of the reviews was low or very low, as the appraisals with the AMSTAR-2 tool found that all reviews had one or more critical flaws. The reviews were largely confined to cohort and case-control studies and focused on more on precision, i.e. quantification of risk, than on causality. Due to the poor quality of the reviews and a lack of information in many reviews regarding the design and quality of included studies, the evidence is largely inconclusive.

The chapter contributes to the literature the first meta-review on the subject of EDCs and breast cancer. This chapter shows that the potential for natural experiments to improve understanding of environmental causes of disease will be hindered if systematic reviews do not include such studies or fail to integrate them into syntheses along with more traditional epidemiological study designs.

This section discusses the implications of the findings for understanding the environmental causes of breast cancer, for the identification and uptake of natural experiments, and for public health research, practice, and policy. The strengths and limitations of the meta-review are then considered.

6.6.1 Interpretation and discussion of evidence on EDCs and breast cancer

The findings of the meta-review are congruent with the change in the IARC status of PCBs from probable to known human carcinogen (Lauby-Secretan et al., 2013) and with ongoing concerns about the toxicity of phthalates (Benjamin et al, 2017). From a public health and a consumer perspective, it is reassuring that there is little evidence of an association between hair dye or deodorant use and increased risk of breast cancer. Furthermore, the reviews did not identify statistically significant increased risks of breast cancer associated with exposure to DDT, dioxin, or cyclodiene pesticides other than heptachlor, which has been banned in many countries (World Health Organization, 2003). However, for many EDCs systematic review evidence is lacking or inconclusive. It is surprising and somewhat disappointing, given the amount of attention devoted to EDCs and the emphasis on related potential health risks, that most reviews were of low quality and unable to support greater certainty.

Although the narrative reviews at least touched on all EDCs of interest, not all EDCs were comprehensively addressed by the included reviews. With reference to the EDC-2 classification of common EDCs (Gore et al., 2015), the risks of breast cancer associated with exposure to atrazine, bisphenols, OCPs (other than DDT), PFOA, and the fungicide vinclozolin were not the subject of comprehensive systematic reviews or meta-analyses. Additionally, this meta-review did not encompass polyaromatic hydrocarbons or endocrine active pharmaceuticals; all of these types of EDCs may warrant further attention and systematic reviews of studies in humans.

In synthesising the evidence and pooling risk estimates, none of the included reviews differentiated between pre- and post-menopausal breast cancer or between hormone receptor positive and negative breast cancers. If EDCs have different places on the causal pathway to different types of breast cancer, which seems plausible, then the reviews and meta-analyses might have restricted their ability to detect this by treating all breast cancers as a single disease. Similarly, as pointed out by Leng et al. (2016) and demonstrated in the phthalate review by Fu et al. (2017), genetic polymorphisms are important modulators on the pathway from environmental exposure to carcinogenesis. It

will be valuable to have more studies that incorporate genetic polymorphisms along with reliable, time-specific assessments of exposure in order to better understand the environmental causes of breast cancer.

6.6.2 The contribution of natural experiments to evidence-based public health

As randomised trials are unlikely to be feasible or ethical on the environmental causes of diseases, including cancer, natural experiments should represent an important and valuable source of evidence in this subject area (Academy of Medical Sciences and Rutter, 2007). Yet across 15 systematic reviews of a relatively prominent public health topic with a substantial evidence base, natural experiments did not feature as a source of knowledge. The presentation of evidence in the reviews did not distinguish study designs according to their ability to address selection and confounding and provided limited information on how exposure and comparison groups were identified, or how and why exposure was thought to differ between groups. Indeed, many natural experimental designs were implicitly excluded from the reviews in the first place when the inclusion criteria specified cohort and case-control studies. These observations suggest that the traditional hierarchy of evidence, which neither assigns a place to natural experiments nor specifically recognises their value, continues to shape systematic reviews in public health.

It could be argued that the absence of natural experiments in these reviews accurately reflects an inability to conduct research on this topic with such designs – that it is not feasible to identify and analyse natural experiments on EDCs. However, some studies could be tentatively identified as potential natural experiments from reviews like Rodgers et al. (2018) that described studies in which exposure was determined by geographic location. Rodgers et al. additionally discussed the possibility of identifying natural experiments where changes in regulations create “distinctive exposure scenarios” (p. 172). Such situations may have already occurred, for example, in changes to Canadian regulations regarding bisphenol A (Government of Canada, 2008). The reviews also identified several episodes of industrial contamination which could potentially be analysed as natural experiments. Unfortunately, such episodes also continue to occur, such as the 2011 phthalate incident in Taiwan, in which

food and drink products were contaminated with phthalates by the manufacturers as an unsafe substitute for emulsifiers (Li et al., 2015, Mitoma et al., 2015).

Even if natural experiments on EDCs and breast cancer were to be identified and analysed, however, they would still face certain challenges given the gaps and limitations identified in this meta-review. Assessing the timing and dose of exposure relative to an individual's development, accurately measuring that exposure in a contemporaneous manner, understanding interactions between and among various chemicals and genetic polymorphisms, all pose challenges for researchers, particularly given the ubiquity of EDCs in the environment and their persistence over time.

6.6.3 Implications for research, practice, and policy

The applicability of this meta-review to public health practice and policy is in some ways limited because of the low quality of the included reviews and the inconclusive nature of much of the evidence. However, a few points relating to research and to public health advice relating to breast cancer risk may be in order.

First, the quality assessment of the included reviews serves as a reminder to researchers of the importance of adhering to the PRISMA statement when conducting and reporting systematic reviews. In particular, reviews need to be based on pre-defined protocols, which ideally would be prospectively registered and publicly available, and the review needs to make explicit reference to this protocol. Providing a list of excluded studies with reasons for exclusion not only is good practice, but is also informative when the review and its findings are compared with other reviews and differences in overlap require explanation. Quality assessment of included studies is crucial and not only should such assessments be conducted, but the results reported and incorporated into the synthesis. In these respects, the findings of this meta-review are supported by a scoping review of the impact of the PRISMA Statement, which found that protocol registration and risk of bias assessment were the poorest performing areas of adherence, with just over 20% of a sample of 2,382 systematic reviews published between 2010 and 2016 based on a registered protocol (Page and

Moher, 2017). Systematic reviews are intended in part to reduce research waste, but arguably a poor-quality systematic review only increases it.

Second, the lack of contribution of natural experiments to this evidence base has implications for systematic reviewers, primary researchers, and funders. This meta-review suggests that the Academy of Medical Sciences advice on recognising the value of natural experiments to identify the environmental causes of disease, issued more than ten years ago, may not yet have had an optimal impact. The Academy's advice to researchers to consider "the relative merits and limitations of different research designs" and "whenever possible, use natural experiments" (p. 13) needs to be taken into account, not only by those conducting primary studies, but also by systematic reviewers. Limiting the inclusion criteria of reviews on environmental causes of disease to cohort and case-control studies unnecessarily excludes the potentially valuable evidence that natural experiments can provide. However, in order to make use of that evidence, reviewers and information scientists will need to become familiar with natural experimental study designs and related terminology; literature search strategies will need to be expanded and search filters amended; quality assessment tools (for both systematic reviews and primary studies) will need to consider the design characteristics of natural experimental studies; and synthesis methods, including assessments of strength of evidence, will need to be developed further.

Finally, even given the caveat that the reviews had serious limitations, the evidence to support assertions that EDC exposure increases risk of breast cancer is less compelling than might be expected given the amount of attention directed towards EDCs by some stakeholders; the position of Cancer Research UK, i.e. that the evidence is poor and inconsistent, is upheld. Although reducing the burden of environmental chemicals is desirable on the precautionary principle as well as from an ecological standpoint, and considerable uncertainty about EDCs persists, based on the findings of this meta-review attention to EDCs as a potential cause of cancer should not detract from a strong focus in public health advice and policy on established, modifiable risk factors for breast cancer, namely body weight, physical activity, breastfeeding, and alcohol consumption (Tamimi et al., 2016; Tamimi, Hankinson, and Lagiou, 2018).

6.6.4 Strengths and limitations of this overview

This meta-review was based on a protocol that was prospectively registered in PROSPERO and is publicly available. The PRISMA statement has been followed as closely as possible, with explicit reporting of comprehensive search strategies, inclusion criteria, excluded studies, and quality assessment of included studies by two independent reviewers. The meta-review is somewhat innovative in its presentation of overlap and its map of evidence, providing new examples of application of emerging methods. The main methodological limitation of this meta-review is that study screening and selection was performed by a single reviewer. It is also possible that, by focusing the search on breast cancer, relevant reviews of all cancers which did not mention breast cancer in the title or abstract were missed. This is likely to be particularly true of reviews of all cancers that had positive findings for other cancers and negative findings for breast cancer. However, the initial retrieval of the search was 4745 records for breast cancer alone; screening records for all cancers would probably not have been feasible. A final limitation is that, as determined at the protocol stage, primary studies cited in the included reviews were not retrieved and re-analysed. Doing so might have provided useful information and increased detection of natural experiments, but was not feasible within the scope of this thesis.

6.7 Chapter summary

This chapter has shown that, despite the putative importance of natural experiments in elucidating environmental causes of disease, such studies do not feature in a sample of systematic reviews on a prominent public health topic, and indeed, if they do exist, were either obscured or missed due to limitations in the reviews and conformity to the established hierarchy of epidemiological evidence. Thus, the chapter supports an argument that there is scope to improve the inclusion and presentation of evidence from natural experiments in public health reviews. The evidence for EDCs as a cause of breast cancer is largely inconclusive, but opportunities to strengthen the evidence base and analyse natural experiments may exist by making use of industrial accidents and geographic information about exposures. In the meantime, prevention efforts should continue to focus on known modifiable risk factors.

7 Discussion

7.1 Chapter overview

This thesis has reported the results of three systematic reviews which provide illustrations of the contribution of natural experiments to answering questions of relevance to public health practice and policy, and examples of how to incorporate natural experimental studies in systematic reviews. The purpose of this chapter is to bring together findings from the three reviews in light of the overall research question and aims of the thesis. First, I consider whether the findings support the assertion that incorporating natural experiments into systematic reviews can provide better evidence to support public health decision making. Second, I consider what changes would need to take place in order for the potential of natural experiments to be more fully realised in public health and what barriers and facilitators exist in relation to these changes. Finally, I discuss the implications of the findings for the conduct and reporting of natural experiments, the conduct of systematic reviews in public health, and for public health knowledge translation as practised by guideline developers and GRADE.

7.2 Summary of findings

7.2.1 Systematic review of regression discontinuity studies

The systematic review of RD studies (chapter 3) contributes the most comprehensive review on the subject to date. This review searched 32 health and social science databases and identified 181 RD studies that investigated the effects of interventions or exposures on health outcomes, more than five times the number of the only other review of RD which used the same inclusion criteria but limited its search to one health database (Moscoe, Bor, and Barnighausen, 2015). The topics covered spanned a broad range of areas of social policy and public health intervention, including air quality, tobacco and alcohol control, early years, health systems, nutrition, and road safety, as well as clinical medicine and epidemiology, showing the wide applicability of RD and natural experimental studies and designs to public health research questions. The analysis of forcing variables and cut-off rules used in the studies provides information that can inform the design of future policy evaluation and help in

the identification of new natural experiments to be analysed. The quality assessment of the studies suggested that overall the design, conduct, and reporting of RD studies can be improved by including a narrative explanation of the implementation of the assignment rule, reporting density and falsification tests, reporting attrition, and pre-specifying a primary outcome.

The strengths of the RD systematic review include following a registered review protocol, conducting an extensive search, double-screening and double-appraisal of a 10% sample of studies, detailed quality assessment of included studies, and reporting according to PRISMA guidelines. The review could have been improved by having all search results, data extraction, and critical appraisal conducted by a second reviewer for all studies rather than a sample, but this was not feasible within the resource limitations of a PhD project. The last date searched was March 2015 and ideally the search could be updated. However, given the number of studies already identified and the fact that the review was not a synthesis of effect estimates, an update search would be unlikely to change the conclusions of the review.

7.2.2 Systematic review of minimum legal drinking age studies

The systematic review of RD studies on MLDA (chapter 4) contributes a new assessment of the effectiveness of this alcohol control intervention and a demonstration of the application of systematic review methods to evidence from natural experiments. Because individuals cannot choose their age or manipulate the legislated threshold, the RD design makes it possible to identify a causal effect of MLDA legislation by comparing outcomes for those just above and below the threshold. The support for causal inference means that this review, which is the first systematic synthesis of RD studies on MLDA, can be argued to present the best available evidence on MLDA and support the assertion that natural experiments can provide useful evidence (in terms of causal inference, external validity, and policy relevance) for decision makers.

The included MLDA studies (n=17) presented several problems for synthesis, resulting in lessons learned that can inform future systematic reviews and thereby help to realise the potential for natural experimental studies to be usefully incorporated into evidence syntheses. Poor and inconsistent reporting of

data would make author contact a likely necessity for conducting meta-analyses, which would require additional resources. These studies report multiple model estimates and sensitivity analyses and generally do not involve a pre-specified model selection method or primary outcome, meaning that selecting an estimate for synthesis is potentially open to bias. The modified effect direction plot was developed in response to this problem and makes it possible to visualise a complex body of natural experimental evidence in one figure. The review concluded on this basis that mortality and hospital admissions probably increase at the MLDA, meaning that the legislation has a protective effect on those below the legal drinking age, but evidence on motor vehicle accidents and drug use is inconsistent. The quality of this evidence was moderate to high based on assessment against the WWC Standards for RD.

The MLDA studies were identified within the larger systematic review of RD reported in chapter 3 and therefore benefit from the registration of a review protocol and the comprehensive search that was conducted. Additional strengths of the MLDA review include quality assessment by two reviewers and an update search (to June 2018) that identified two new studies, neither of which would change the conclusions of the review. Limitations of the MLDA review include data extraction that was performed by only one person, lack of time and resources to contact study authors for additional data that might have made meta-analysis possible for some outcomes, and an inability to assess publication bias due to the presentation of multiple model estimates within studies.

7.2.3 Critical appraisal checklist for RD studies

A further product of the RD systematic review was the development of a critical appraisal checklist for RD studies (chapter 5). Tools specific to natural experimental designs will need to be developed in order to incorporate such studies into evidence syntheses and ensure that the results of natural experiments can be evaluated and used to support decision-making. This chapter contributes one such tool which benefited from testing with three users and application to 181 studies. However, the checklist is at an early stage of development and requires further testing and refinement to ensure good interrater reliability.

7.2.4 Meta-review on endocrine-disrupting chemicals and breast cancer

The overview of systematic reviews on EDCs and breast cancer (chapter 6) contributes the first meta-review on this subject. Fifteen systematic reviews published between 2005-2018 were identified including eleven meta-analyses. Overall there was no statistically significant increase in the pooled relative risk of breast cancer associated with total exposure to any of the classes of EDCs examined, although there was a statistically significant increased risk for exposure to some congeners of PCBs and in some subgroup analyses for phthalates. The quality of the included reviews was low or critically low according to assessment with the AMSTAR-2 tool, chiefly owing to failure to report that the review was based on a protocol. A map of the evidence identified some strong evidence for PCBs, phthalates, and household or consumer products, but for all other EDCs covered in the reviews the evidence was inconclusive or weak.

In terms of assessing the contribution of natural experiments to the evidence base on EDCs and breast cancer, the findings of the meta-review were negative. None of the reviews identified any included studies as natural experiments; ten of the 15 reviews limited the inclusion criteria to cohort and case-control or ‘epidemiological’ studies. However, the narrative reviews provided some details suggestive of natural experiments in descriptions of the exposure mechanisms of some studies, which related to industrial accidents or geographical assessments of exposure. The decision not to retrieve and re-examine the primary studies included in the reviews was made at the protocol stage and was appropriate given the project scope and the large number of primary studies included in the reviews, but does constitute a limitation in the ability to identify the contribution of natural experiments, which was not foreseen at the protocol stage.

Additional limitations of the meta-review include study selection and data extraction performed by a single reviewer. Furthermore, it is possible that negative findings were missed by limiting the inclusion criteria to reviews of breast cancer risk and excluding systematic reviews that examined ‘all’ cancers or health outcomes with no mention of breast cancer. However, the findings of

included reviews were almost entirely negative in terms of finding a statistically significant increased risk. The strengths of the meta-review include a registered protocol developed according to the PRISMA-P standard, a comprehensive search of six databases and grey literature, and quality assessment by two reviewers using a validated tool.

7.3 Implications for conduct and reporting of natural experiments

The experience of extracting data for synthesis (chapter 4) and the results of the quality assessment (chapter 5) both support the conclusion that reporting of RD studies needs improvement in order to allow systematic reviewers and other users to assess whether design assumptions and quality criteria have been met and to ensure that study results are understood and correctly interpreted. Clear reporting of how the threshold rule was implemented is essential in order to show that the RD design was valid, that treatment allocation was free of selection bias, and that the treatment effect unconfounded. Reports of RD studies should also include evidence from appropriate tests of the underlying design assumptions, namely density and falsification tests. The latter should include tests for spurious discontinuities at the cut-off for outcomes that ought not to be affected by the intervention (such as hospital admissions for appendicitis at the MLDA) and at values of the forcing variable other than the cut-off (such as age 23 for an MLDA of 21).

The thesis supports the need for improved reporting of RD studies, but as similar findings have been reported for IV (Davies et al., 2013) and ITS (Ramsay et al., 2003), and most if not all analyses of natural experiments depend upon some underlying assumptions and some kind of modelling, it seems reasonable to conclude that reporting quality is an issue that deserves consideration across natural experimental designs. The MRC guidance on natural experiments also noted the importance of transparent reporting and the need to follow guidelines such as TREND or STROBE. Although many STROBE checklist items are generic and applicable to any non-randomised study, others are specific to cohort, case-control, or cross-sectional studies. Extensions of STROBE have been prepared to address specific areas such as nutritional epidemiology, genetic association studies, and studies based on routinely collected healthcare data (vol Elm et al.,

2007, Lachat et al., 2016, Little et al., 2009, Benchimol et al., 2015). An extension of STROBE for natural experimental studies might provide an effective tool to promote better reporting, incorporating criteria already identified in the MRC guidance and elsewhere (Craig et al., 2012, Craig et al., 2017, Lee and Lemieux, Dunning).

Development of reporting standards for natural experiments could be beneficial in raising awareness of these designs and disseminating knowledge about good practice in study design, conduct, and reporting. However, evidence as to whether the publication of reporting standards actually has a positive effect on reporting quality is mixed (Page and Moher, 2017). In order to increase the likelihood that the effort put into producing such standards resulted in the desired improvement in reporting, any such standards should be accompanied by dissemination and impact plans and interventions to increase adherence. Ideally, the effectiveness of the standards could be evaluated in a prospective controlled study.

7.4 Implications for systematic reviews

Given the potential demonstrated in the thesis for RD studies and, by extension, natural experimental studies, to provide relevant and useful evidence for public health, it seems reasonable to conclude that some revision to systematic review methods and development of related tools should be considered in order to ensure that such studies are identified and incorporated into public health reviews. The examples of the application of systematic review methods in this thesis, as well as the findings of the meta-review, support several suggestions of changes to methods used by Cochrane, HTA and guideline development organisations that conduct systematic reviews to inform health system decisions, and GRADE. These changes relate to inclusion criteria, literature searching, risk of bias assessment, data extraction, and methods for synthesising these studies; they also need to be considered during protocol development as well as during the execution of the review.

In terms of the capacity of systematic reviews to enable the uptake of natural experimental studies as evidence, this thesis has demonstrated limitations within a sample of systematic reviews in terms of their ability to identify or

describe in detail the context and findings from these study types. Reviews on questions amenable to investigation through natural experiments, such as those addressing environmental causes of disease or evaluation of population-level interventions and policies, should not be limited to cohort and case-control studies and need to search a range of social science as well as health databases. However, the chapters on RD and MLDA also demonstrate the utility of systematic review methodology in demonstrating how natural experimental designs can be applied to a wide variety of public health research questions, although some adaptation may be required for critical appraisal, data extraction, and synthesis. A comprehensive systematic review in public health should be designed at the protocol stage to consider the potential relevance of natural experiments to the research question and specify inclusion and exclusion criteria, search strategies, quality assessment, and synthesis plans accordingly.

7.5 Implications for knowledge translation

The findings of the thesis that are relevant for systematic review methods by extension have implications for developers of evidence-based guidelines, which use or adapt such methods. Guidelines are an important knowledge translation product through which evidence may inform public health practice, health service organisation, and health policy. Evidence-based guideline development is guided by methodologies which vary in their prescriptiveness. A methods manual such as NICE PMG20 is sufficiently flexible to support the incorporation of natural experimental studies as evidence (National Institute for Health and Care Excellence, 2015); however, anecdotal evidence from colleagues at NICE suggests that a lack of familiarity with natural experimental designs is a barrier to their inclusion. Other guideline development processes may take a more restrictive approach to evidence which may inadvertently prevent the uptake of natural experimental studies. For example, the SIGN 50 handbook (Scottish Intercollegiate Guidelines Network, 2015) describes the use of design-specific search filters and appraisal methods that might act as a barrier to including natural experimental designs such as RD, DiD, IV, or synthetic controls, for which such tools are lacking.

The examples of RD evidence identified in this thesis may be of use in developing GRADE guideline methods for application in public health. The GRADE

approach has been developed to promote transparency in guideline methodology and reduce unnecessary variation in methods. Several studies have reported challenges in developing public health guidelines using GRADE (Akl et al., 2012, Alexander et al., 2016, Rehfuess and Akl, 2013), with the treatment of non-randomised studies and the strength of recommendations frequently cited as a source of concern. Natural experimental studies may provide examples of high certainty (without upgrading) in non-randomised evidence that are currently lacking in the GRADE literature (Schünemann et al., 2018) and thereby demonstrate another way in which strong recommendations can be supported in evidence-based public health guidelines.

7.6 Implications for evidence-based public health

The implications identified above (sections 7.3 to 7.5) for the reporting, synthesis, and translation of evidence suggest action is needed to support better reporting of natural experiments, to ensure that they are included in public health systematic reviews, and to promote their translation into public health policy. The underlying assumption of benefit from these outcomes is that better research evidence and syntheses will support the implementation of interventions and policy that will in turn result in better health for the public. This assumption is simplistic and needs to be tempered by knowledge of the barriers and facilitators to evidence-informed policy making (Armstrong et al., 2014; Ellen et al., 2014) and by the recognition that research is only one, not necessarily privileged, source of information and ideas that influence policy decisions (Smith, 2013). However, there are further benefits to be realised from making these changes to the tools and methods of evidence-based public health that go beyond the production of better evidence and syntheses to support decision-making.

These benefits include breaking down disciplinary silos and increasing the development of novel and interdisciplinary approaches to public health problems, which Hanlon et al. (2012) argue is necessary in order to meet future public health challenges. An additional benefit is an increased potential for research to be designed and funded to investigate interventions and approaches not amenable to randomisation, particularly with regard to investigations of equity and transferability across contexts (Waters, 2009), so that actions on

populations, social determinants and health inequalities are not less likely than actions on a sample of individuals to be supported by strong recommendations.

Consider, for example, the WHO Guidelines on Integrated Care for Older People (World Health Organization, 2017), categorised by the WHO as a guideline for health systems. The challenges of providing integrated care for older people from a systems perspective include co-ordination of health and social service provision, integrated access to medical and social care records, provision of welfare benefits and insurance coverage, ensuring safe transitions between home and care settings, and significant funding challenges for the health system. Evidence to inform approaches to these challenges could well be provided by natural experimental studies. Yet the WHO guideline's 13 recommendations all relate to interventions delivered to individuals and largely evaluated with RCTs: multimodal exercise, oral nutrition supplements, cognitive stimulation, and fall prevention. Given that interventions aimed at individuals may not produce as great an effect on population health as a population approach (Rose 1981, Rose 2001), the conduct, funding, and dissemination of natural experimental studies could help to ensure that systems- and population-level interventions receive better evaluations and more recognition in evidence-based public health guidelines, with resulting benefits for population health.

Even though the hierarchy of evidence has been repeatedly challenged and revisions proposed, its influence can still be seen in evidence synthesis methodologies including Cochrane and GRADE. This influence is seen in ongoing debates as to whether and how to include NRS and RS in the same review, the approach taken in the ROBINS-I NRS critical appraisal tool of using an imaginary randomised trial as a starting point for assessment, and the GRADE approach to strength of evidence in which randomised trials start as high quality and NRS start as low. It may be timely to ask whether randomisation as a surrogate for 'strength of evidence' should be re-examined and to articulate what it is that randomised designs achieve in terms of causal inference, and under what assumptions (Deaton and Cartwright, 2018; Gelman, 2018; Cook, 2018).

In the first instance one could suggest that elements such as unconfoundedness, absence of selection effects, and testing and rejection of alternative hypotheses should have a more prominent role in judging the certainty or strength of

evidence to inform decision-making. Additionally, the potential exists for a combined approach to risk of bias assessment of RS and NRS if randomisation is 'unpacked' and assessment focuses on these elements and whether design assumptions have been met. A further consideration relates to how evidence-based or evidence-informed approaches can more meaningfully and usefully synthesise the broad range of sources of evidence that are relevant to decision-making. Ultimately the aim of such developments would be to help achieve the goal towards which this thesis has also been directed: the inclusion of a wider range of study designs that will enable the production of systematic reviews of greater trustworthiness, relevance, and utility to decision making, which in turn support actions of greater benefit to the public health.

7.7 Recommendations for research and methodological development

This thesis concludes by translating the findings (summarised in section 7.2) and their implications (sections 7.3 to 7.6) into a set of recommendations. These recommendations address three areas of public health research and practice: (1) further research relating to RD and other natural experimental study designs, (2) additional systematic reviews and related methods research, (3) actions guideline developers and others involved in public health knowledge translation may take to ensure uptake of natural experimental studies.

7.7.1 Recommendations for further research: RD and other natural experimental study designs

- Investigation of differences in design, assumptions, and estimates between date-based RD and ITS in order to guide the choice of design for analysis of natural experiments in which exposure or treatment allocation has a time element.

- Development and dissemination of reporting standards for RD specifically and natural experimental studies generally, with prospective evaluation of the impact of the standards.

- Classification of sources of risk of bias in natural experimental study designs, with a view to considering whether design assumptions and

statistical assumptions should also be part of the assessment of study quality.

- Quantification of risk of bias in natural experimental study designs in order to support the development of critical appraisal tools and reporting standards with empirical evidence.

- Replication of RD studies in different contexts using different data sources, applying the findings relating to common types of forcing variables and cut-off rules to aid in identifying new natural experiments.

7.7.2 Recommendations for further research: systematic reviews and related methods

- Methods-based systematic reviews of the application and reporting practice of natural experimental designs in addition to RD (such as ITS and DiD) in public health in order to assess their quality and identify challenges and solutions in synthesising such evidence.

- Topic-based systematic reviews of common natural experiment scenarios (such as natural disasters and legislative changes) in public health in order to identify best practices in design and reporting as well as opportunities for further replication.

- Comparison of findings from studies that use different methods to evaluate the same natural experiment (either within or across studies), in order to determine any association between method and effect estimate and to inform which estimates should be extracted for synthesis in systematic reviews. For example, how findings differ if a similar research question is investigated using difference-in-differences, RD, and ITS analyses.

- Development and user testing of critical appraisal and data extraction tools for natural experimental studies.

- Development and user testing of the effect direction plot and other novel visualisation methods to represent studies which report a range of estimates for each outcome.

- Investigation and development of guidance on translation of outcomes from econometric studies into common metrics to aid quantitative synthesis.

- Development of guidance on incorporating natural experimental studies in systematic reviews.

7.7.3 Recommendations for guideline developers

- Ensure that methodologists and other technical staff have sufficient training for their role to identify, appraise, and synthesise results from natural experimental studies.

- Consider providing more detailed methodological guidance and examples to facilitate the incorporation of natural experimental studies into literature searches, summary of findings tables, and evidence synthesis.

- Inform and participate in the development of new critical appraisal checklists and other evidence synthesis tools to ensure their usability and subsequent uptake.

- Incorporate considerations of causal inference when applying GRADE to questions that use nonrandomised evidence, allowing panel members to articulate how the evidence base supports causal inference and potentially allowing for stronger recommendations in public health and policy.

Appendices

Appendix 1	Protocol: Evaluation of public health interventions using regression discontinuity designs: a systematic review [CRD42015025117]
Appendix 2	Detailed characteristics of included studies for chapter 3
Appendix 3	Detailed critical appraisal results for chapter 4
Appendix 4	Protocol: Endocrine disrupting chemicals and the risk of breast cancer: a systematic review of reviews [CRD42018089344]
Appendix 5	Literature search strategies
Appendix 6	Detailed critical appraisal results for chapter 6

Appendix 1. Protocol: Evaluation of public health interventions using regression discontinuity designs: a systematic review [CRD42015025117]

Evaluation of public health interventions using regression discontinuity designs: a systematic review

Michele Hilton Boon, Peter Craig, Laurence Moore, Hilary Thomson

Citation

Michele Hilton Boon, Peter Craig, Laurence Moore, Hilary Thomson. Evaluation of public health interventions using regression discontinuity designs: a systematic review. PROSPERO 2015 CRD42015025117 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42015025117

Review question

1. How and in what areas of research have regression discontinuity designs been applied to evaluate the health impacts of public health interventions and policy?

2. What is the quality of reporting in studies using regression discontinuity designs to evaluate health-related outcomes?

Searches

Language: all languages will be included.

Dates: since 1960 (date of first publication describing regression discontinuity methods).

Search strategy:

Relevant studies will be identified using the search term “regression discontinuity” (title, abstract, keyword) and any equivalent subject index terms. Regression discontinuity designs are more commonly used in social sciences (particularly economics, education, and political science) than in health research; accordingly, the search strategy will include databases from these fields in addition to health databases.

The databases to be searched are: ASSIA, Business Source Premier, CINAHL, Cochrane Library, EBSCO Professional Development Collection, EconLit, EMBASE, ERIC, EThOS (British Library Electronic Theses Online Service), Google Scholar, IDOX, International Bibliography of the Social Sciences, King’s Fund Publications, MEDLINE (PubMed), MEDLINE In Process, NICE Evidence Search, NTIS, Open Grey, POPLINE, ProQuest Dissertations and Theses Database, PsycINFO, RePeC, Scopus, Social Care Online, Social Services Abstracts, SocINDEX, Sociological Abstracts, TRIP, US Environmental Protection Agency document repository, Web of Science, WHO Institutional Repository, World Bank Documents & Reports.

Types of study to be included

Included studies must use a regression discontinuity design. Included studies may also use additional designs, such as difference-in-difference.

Condition or domain being studied

All public health policy areas, including but not limited to: air quality, alcohol and substance misuse, early years interventions, food policy, nutrition and obesity, maternal and infant health, mental health promotion and suicide prevention, health service organisation and delivery, housing, transportation, tobacco, sexual health, screening, vaccination programmes.

Participants/population

Any populations whose eligibility for a public health intervention or programme is determined by a cut-off value of a continuous variable, making evaluation of the programme suitable for a regression discontinuity design.

Intervention(s), exposure(s)

Any public health intervention, programme, or policy involving treatment assignment based on a cut-off rule, including but not limited to: age (such as minimum legal drinking age or vaccination schedule), income (such as early years or housing improvement programmes), time (such as imposition of a legislative ban), a clinical score or biological variable (such as birthweight).

Comparator(s)/control

Non-exposed control group (below cut-off value of forcing variable).

Context

The regression discontinuity (RD) design was first proposed by Thistlethwaite and Campbell (1960) based on the intuition that, given an eligibility rule based on a cut-off value for a continuous variable whose value cannot be precisely manipulated by participants or administrators, treatment assignment will be effectively random within a certain bandwidth on either side of the cut-off and therefore, differences in an outcome affected by the treatment can be estimated as the difference between groups just above and just below the cut-off, without any bias due to unobservables.

Moscoe, Bor, and Barnighausen (2015) argue that RD is likely to be useful in health research because the use of cut-off rules for treatment assignment is common. Their review identified 32 studies from medicine, epidemiology, or public health that used an RD design; however, their search was restricted to a single database (PubMed).

We aim to conduct a systematic review that draws on a variety of disciplines and sources to determine how RD designs have been used to analyse the health effects of natural experiments in the wide range of policy areas relevant to public health.

Primary outcome(s)

1. Direction of effect in any health-related outcome, e.g. hospital admissions, mortality.

2. Forcing variables, cut-off values, bandwidth, and analytical methods used.

Secondary outcome(s)

Assessment of study quality.

Data extraction (selection and coding)

One author will perform an initial screen of titles and abstracts, remove duplicates, and exclude studies that clearly do not meet eligibility criteria. A second author will screen a random sample of studies (10%) to verify that eligibility criteria have been consistently and correctly applied. Two authors will independently review the full text of the remaining papers and determine eligibility. EndNote (version X7) will be used to record reasons for exclusion. Disagreement will be resolved by discussion and consensus or, when this is not achieved, by having a third author review the paper for eligibility. Where insufficient information is provided in the paper to make a decision about eligibility or to complete data extraction, one attempt to contact the study authors will be made.

Information will be extracted from each included study relating to: citation details (author, date, country); the population under investigation; the intervention, event, or change under investigation; the control or comparison group; the forcing variable and cut-off used; the health outcome(s) reported; the statistical methods used; the main findings; and study quality/risk of bias.

Risk of bias (quality) assessment

Two authors will independently assess the quality of each included study using the Standards for Regression Discontinuity Designs produced by the What Works Clearinghouse. Disagreements will be resolved by discussion and consensus or, when this is not achieved, by having a third author complete an additional assessment. Results will be presented in tabular and graphical formats to provide an overview of study quality.

Strategy for data synthesis

This review is designed to integrate studies that address a wide variety of research questions from different disciplines and policy areas. Accordingly, synthesis methods will be developed iteratively from an initial configurative mapping of the literature and tabulation of study characteristics. Extracted data will be presented in tables to describe RD design elements, estimates of effect, and study quality. Results will be presented by type of intervention and by policy area in order to enable readers to identify applications of RD in areas most relevant to their research interests. Graphs of the number of studies by year and by discipline will enable identification of trends in the use of RD. As the review is not designed to identify studies that answer a particular research question, no meta-analysis is planned. If, however, several studies do happen to answer similar questions, forest plots will be used to demonstrate how estimates of effect sizes and directions of effect differ across studies.

Analysis of subgroups or subsets

More detailed analysis of subsets of studies will be conducted if multiple studies are identified that evaluate the same intervention or that investigate sufficiently similar policy questions.

Contact details for further information

Michele Hilton Boon

m.boon@sphsu.mrc.ac.uk

Organisational affiliation of the review

MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

http://www.sphsu.mrc.ac.uk/

Review team members and their organisational affiliations

Ms Michele Hilton Boon. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Dr Peter Craig. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Professor Laurence Moore. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Dr Hilary Thomson. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

Anticipated or actual start date

05 January 2015

Anticipated completion date

02 December 2015

Funding sources/sponsors

Medical Research Council doctoral studentship

Conflicts of interest

None known

Language

English

Country

Scotland

Stage of review

Review_Ongoing

Subject index terms status

Subject indexing assigned by CRD

Subject index terms

Humans; Health Services; Public Health

Date of registration in PROSPERO

04 August 2015

Date of publication of this version

04 August 2015

Appendix 2. Characteristics of regression discontinuity studies of health outcomes

Table A1.1. Regression discontinuity applications in the evaluation of population-level interventions, by public health policy area.

Study	Context	Forcing variable	Intervention	Outcome(s)
Air quality (5 studies)				
Chay, K. Y. and M. Greenstone. (2003)	USA	total suspended particulates, TSPs (air pollution regulatory threshold)	Clean Air Act Amendments (1970)	Infant mortality
Neidell, M. (2010)	USA	ozone forecast threshold rule for issuing smog alerts	Smog alerts	Outdoor activities (attendance at outdoor venues)
Noonan, D. S. (2014)	USA	Ozone forecast level	Air quality alerts	Amount and intensity of outdoor activity; driving
Sanders, N. J. and C. Stoecker (2015)	USA	total suspended particulates, TSPs (air pollution regulatory threshold)	Clean Air Act Amendments (1970)	sex ratio of live births as estimate of averted fetal losses
Yang, M. (2008)	USA	total suspended particulates, TSPs (air pollution regulatory threshold)	Clean Air Act Amendments (1970)	Infant mortality
Alcohol and substance abuse (18 studies)				
Boes, S. and S. Stillman (2013)	New Zealand	Age	Decrease in minimum legal	Alcohol consumption

Study	Context	Forcing variable	Intervention	Outcome(s)
			drinking age (MLDA)	
Callaghan, R. C., J. M. Gatley, M. Sanches and M. Asbridge (2014) American Journal of Preventive Medicine	Canada	Age	MLDA	motor vehicle collisions
Callaghan, R. C., M. Sanches and J. M. Gatley (2013). Addiction	Canada	Age	MLDA	alcohol-related hospital events
Callaghan, R. C., M. Sanches, J. M. Gatley and J. K. Cunningham (2013)	Canada	Age	MLDA	alcohol-related hospital events
Callaghan, R. C., M. Sanches, J. M. Gatley and T. Stockwell (2014) Drug and Alcohol Dependence	Canada	Age	MLDA	Mortality - all causes, external causes, MVA
Carpenter, C. and C. Dobkin (2009)	USA	Age	MLDA	Mortality (all causes, external causes, internal causes)
Carpenter, C. and C. Dobkin (2011)	USA	Age	MLDA	Mortality, alcohol consumption

Study	Context	Forcing variable	Intervention	Outcome(s)
Carpenter, C., Dobkin, C. and C. Warman (2014)	Canada	Age	MLDA	Alcohol consumption, mortality (all causes, external causes, internal causes, motor vehicle accidents, injuries)
Carpenter, C. and C. Dobkin (2015)	USA	Age	MLDA	ED visits and inpatient hospitalisations
Conover, E. and D. Scrimgeour (2013)	New Zealand	age and date	Decrease in MLDA	alcohol-related hospital admissions
Crost, B. and S. Guerrero (2012)	USA	Age	MLDA	alcohol consumption
Crost, B. and D. I. Rees (2013)	USA	Age	MLDA	marijuana consumption
Deza, M. (2015)	USA	Age	Alcohol consumption	Consumption of hard drugs
Ertan Yoruk, C. and B. K. Yoruk (2015)	USA	Age	MLDA	Risky sexual behaviour
Ertan Yörük, C. and B. K. Yörük (2012)	USA	Age	MLDA	Psychological wellbeing
Lindo, J. M., P. Siminski and O. Yerokhin (2014)	Australia	Age	MLDA	MVAs, hospitalizations, drinking behaviour
Yörük, B. K. and C. E. Yörük (2011)	USA	Age	MLDA	Alcohol consumption, smoking, marijuana use

Study	Context	Forcing variable	Intervention	Outcome(s)
Yu, B. and D. T. Kaffine (2011)	USA	Date of policy change	Increased alcohol availability following repeal of Sunday alcohol sales restriction	Alcohol-related traffic accidents and traffic citations
Disease prevention and screening (5 studies)				
Kadiyala, S. and E. Strumpf (2011)	USA	Age	guideline recommendations regarding age for asymptomatic screening	Marginal benefits of breast, colorectal and prostate screening
Kadiyala, S. and E. C. Strumpf (2011)	USA and Canada	age	guideline recommendations regarding age for asymptomatic screening	tests for breast, colorectal, and prostate cancers
Rashad, H. (1992)	Egypt	Year of programme implementation	National Control of Diarrheal Diseases Project	Infant mortality
Smith, L. M., J. S. Kaufman, E. C. Strumpf and L. E. Levesque (2015)	Canada	Quarter of birth	HPV vaccination	Composite indicator of sexual behaviour
Ziegelhöfer, Z. (2012)	Guinea	Investment cost per inhabitant (programme eligibility criterion)	Rural water supply and hygiene education programme	Prevalence of diarrhoeal disease in children under 5 years
Early years (9 studies)				

Study	Context	Forcing variable	Intervention	Outcome(s)
Carneiro, P. and R. Ginja (2014)	USA	Family income (programme eligibility cutoff)	Head Start	Health measures from CNLSY
Coburn, J. L. (2009)	USA	age of child on 30 Sept 2007	Prekindergarten programme	Brigance Screen age-equivalent scores
Gormley, W.T., Gayer, T., Phillips, D. & Dawson, B. (2005)	USA	Birthdate	Oklahoma universal prekindergarten program	School readiness, as measured by three subtests of Woodcock-Johnson Achievement test)
Lipsey, M. W., D. C. Farran, C. Bilbrey, et al. (2011)	USA	Birthdate	Tennessee voluntary pre-kindergarten programme	School readiness (Woodcock Johnson III test)
Ludwig, J. and D. L. Miller (2007)	USA	County poverty rate	Head Start	mortality rate
Rosero, J. and H. Oosterbeek (2011)	Ecuador	Programme proposal quality score (assigned by funding body)	Early childhood programmes (home visits and childcare centres) for poor families)	Multiple child health and development measures; maternal stress and depression
Santos, R. G. (2006)	Canada	Family Stress Checklist score (programme eligibility rule)	BabyFirst home visit programme	Family social support, parental mental health, parenting outcomes
Weiland, C. and H. Yoshikawa (2013)	USA	Birthdate (programme eligibility cut-off)	Boston Public Schools prekindergarten programme	Cognitive, executive function and emotional

Study	Context	Forcing variable	Intervention	Outcome(s)
				development outcomes
Wong, V.C., Cook, T.D., Barnett, W.S. & Jung, K. (2008)	USA	Birthdate	State pre-kindergarten programmes	Children's cognitive skills/school readiness (receptive vocabulary, math, print awareness)
Healthcare organisations and systems (10 studies)				
Almond, D. and J. J. Doyle Jr (2011)	USA	Time of birth (minutes from midnight)	length of stay / minimum LOS legislation	hospital readmissions and infant mortality
Coudin, E., A. Pla and A.-L. Samson (2014)	France	Year (that GP commenced practice)	Reform of GP billing regulations	GP care provision, fees, prescribing behaviour
Daysal, N. M., M. Trandafir and R. Van Ewijk (2013)	Netherlands	weeks of gestation (week-37 referral rule)	Obstetrician supervision of preterm birth	7- and 28-day mortality, Apgar score
Del Bono, E., M. Francesconi and N. G. Best (2011)	UK	date health warning issued	UK Committee on Safety of Medicines health warning re combined oral contraceptives and risk of VTE	daily average numbers of conceptions, abortions, and live births; neonatal health outcomes ("quality of birth")
Glance, L. G., T. M. Osler, D. B. Mukamel, J. W. Meredith and A. W. Dick (2014)	USA	Date of intervention	Performance benchmarking (nonpublic hospital report cards)	in-hospital mortality

Study	Context	Forcing variable	Intervention	Outcome(s)
Koch, S. F. (2013)	South Africa	Age	Policy change in fees for public healthcare	Healthcare utilisation
Marier, A. (2014)	USA	Proportion of low-income patients	Medicare DSH (Disproportionate Share Hospital) status	Patient experience scores
Sojourner, A. J., R. J. Town, D. C. Grabowski and M. M. Chen (2012)	USA	Unionisation vote share	Unionisation in nursing homes	Care quality (based on state inspection data)
Williams, S. V. (1990)	USA	Year	cost-monitoring letters to physicians from insurer	mean of total billed charges per year
Zuckerman, I. H., E. Lee, A. K. Wutoh, Z. Xue and B. Stuart (2006)	USA	Number of monthly SAB inhaler prescriptions	Drug utilisation review letter to prescribers	Change in monthly SAB inhaler prescriptions
Nutrition and obesity (6 studies)				
Capacci, S., M. Mazzocchi and B. Shankar (2012)	France	Age	Vending machine ban	Calorie and nutrient intakes reported in national nutritional surveys (7-day food diary)
Meller, M. and S. Litschig (2014)	Ecuador	Poverty index (programme eligibility criterion)	PANN2000 food supplementation and health check programme	Child mortality, fertility
Olsho, L. E. W., J. A. Klerman, L.	USA	proportion of students eligible for	US Dept of Agriculture Fresh	24-hour dietary intake

Study	Context	Forcing variable	Intervention	Outcome(s)
Ritchie, P. Wakimoto, K. L. Webb and S. Bartlett (2015)		free or reduced-price meals (state-specific program funding cut-off)	Fruit and Vegetable Program	
Peckham, J. G. and J. D. Kropp (2012)	USA	Family income to poverty ratio	National School Lunch Program	Obesity (BMI, WHR, %body fat)
Schanzenbach, D. (2009)	USA	Income to poverty ratio	National School Lunch Program	Child obesity
Road safety (3 studies)				
Burger, N. E., D. T. Kaffine and B. Yu (2014)	USA	Time (date of ban)	Legislative ban on handheld cell phone use while driving	Number of daily traffic accidents
De Paola, M., V. Scoppa and M. Falcone (2013)	Italy	Date legislation introduced	Penalty points system for traffic offences	Traffic accidents, injuries, and fatalities
Hansen, B. (2015)	USA	Blood alcohol content	BAC-based punishments for drink-driving	Recidivism
Tobacco (5 studies)				
Pieroni, L., M. Chiavarini, L. Minelli and L. Salmasi (2013)	Italy	Year of smoking ban	Indoor smoking ban	Quitting, cigarette consumption, alcohol consumption
Pieroni, L. and L. Salmasi (2015)	Italy	Year of smoking ban	Indoor smoking ban	BMI
Waller, B. J., J. E. Cohen, R. Ferrence, S. Bull and E. M. Adlaf (2003)	Canada	Year	Decrease in cigarette prices	Youth smoking prevalence and mean cigarettes smoked per day

Study	Context	Forcing variable	Intervention	Outcome(s)
Yan, J. (2014)	USA	Maternal age at conception	Minimum cigarette purchase age	Prenatal smoking, infant health measures
Yoruk, C. E. and B. K. Yoruk (2014)	USA	Age	Minimum legal tobacco purchase age laws	Smoking behaviours

Table A1.2. Regression discontinuity applications in medical and nursing interventions (excluding psychiatry)

Study	Context	Forcing variable	Intervention	Outcome(s)
Almond, D., J. J. Doyle, Jr., A. E. Kowalski and H. Williams (2010)	USA	Birthweight (VLBW threshold of 1500g)	Medical care for VLBW infants	Mortality and hospital costs
Bharadwaj, P., K. V. Løken and C. Neilson (2012)	Norway and Chile	Very low birthweight and date surfactant therapy introduced	Extra medical attention and lung surfactant therapy	Mortality and academic achievement
Bor, J., E. Moscoe, P. Mutevedzi, M. L. Newell and T. Barnighausen (2014)	South Africa	CD4 count	ART for HIV	mortality hazard
DISMEVAL Consortium (2012)	Spain	CV risk score	Nurse-led structured telephone interview on CVD risk and prevention	Cholesterol, BP, BMI, CV risk score
Garrouste, C., J. Le and E. Maurin (2011)	France	Risk score for Down syndrome	Reimbursement eligibility	Amniocentesis and foetal health
Jensen, V. M. and M. Wust (2015)	Denmark	Date (of information shock in form of	Caesarean section for breech births	APGAR score, GP visits, severe

		early RCT publication)		morbidity, hospitalizations, complications, infections
Sloan, F. A. and B. W. Hanrahan (2014)	USA	Year	Introduction of photodynamic therapy and anti-VEGF therapies for ARMD	Vision loss or blindness, depression, admission to long-term care facility
Zhao, M., Y. Konishi and P. Glewwe (2013)	China	Systolic blood pressure	Hypertension diagnosis	Dietary intake (fat, protein, carbohydrates, energy) and use of anti-hypertensive drugs

Table A1.3. Regression discontinuity applications in psychology and psychiatry

Study	Context	Forcing variable	Intervention or Exposure	Outcome(s)
Høglend et al. (1993)	Norway	Score based on selection criteria for psychotherapy	"transference interpretations" within brief dynamic psychotherapy	Changes in various clinical assessment scales
CATS Consortium (2010)	USA	trauma score (PTSD Reaction Index)	trauma-specific CBT	6-month change in trauma score
Daniels, V., M. Somers, J. Orford and B. Kirby (1992)	UK	Exeter Alcohol Scale (pre-intervention)	Advice and self-help manual on reducing alcohol consumption	Exeter Alcohol Scale (post-intervention)

Study	Context	Forcing variable	Intervention or Exposure	Outcome(s)
Devitt, T. S. (2006)	USA	Date (of policy change)	Rescinding zero-tolerance policy for onsite substance abuse in a residential treatment centre	Substance Abuse Treatment Scale, breathalyser and urine toxicology screening
Elder, T. E. (2010)	USA	Birthdate (relative to state kindergarten eligibility cutoff)	School starting age	ADHD symptoms, diagnosis and treatment
Evans, M. E., S. M. Banks, S. Huz and T. L. McNulty (1994)	USA	Date of intervention	Intensive case management programme	State psychiatric hospital use
Evans, W. N., M. S. Morrill and S. T. Parente (2010)	USA	Birthdate (relative to state kindergarten eligibility cutoff)	School starting age	ADHD diagnosis and treatment
Flam-Zalcman, R., R. E. Mann, G. Stoduto, et al. (2013)	Canada	addiction severity measure	Alcohol brief intervention programme	Alcohol use
Høglend, P. (1996)	Norway	pretest suitability measure	brief dynamic psychotherapy	"overall dynamic change"
McFarlane, W. R., B. Levin, L. Travis, et al.	USA	Positive Symptoms Scale	FACT (Family-aided Assertive Community Treatment) package	conversion to psychosis, as defined by positive symptoms
Mezuk, B., G. L. Larkin, M. R. Prescott, et al. (2009)	USA	Date (11 September2001)	11 September 2001 terrorist attacks	monthly suicide rate per 100,000 in NYC

Study	Context	Forcing variable	Intervention or Exposure	Outcome(s)
Pesko, M. F. (2014)	USA	Dates of terrorist attacks	Terrorist attacks	Stress, smoking
Yang, M. (undated)	USA	Date	September 11th terror attack-induced anxiety	Marijuana use

Table A1.4. Regression discontinuity applications in the investigation of health outcomes of social policies in developed countries. These studies investigate the indirect health effects of policy interventions and programmes that were not specifically designed or intended to effect a change in the specified health outcome at the population level (see Table 3) or the individual level (see Tables 4 and 5).

Study	Context	Forcing variable	Intervention	Outcome(s)
Beuchert, L. V., M. K. Humlum and R. Vejlin (2014)	Denmark	Date of reform	Reform of maternity leave laws	Hospital admissions, ED visits, maternal depression, family outcomes
Boheim, R. and T. Leoni (2014)	Austria	Firm's wage sum (threshold for paying deductible on sickness absence insurance)	Deductible of 30% payable by large employer on sickness absence insurance	Blue-collar workers' sickness absences
Garcia-Gomez, P. and A. C. Gielen (2014)	Netherlands	Age (45, threshold for exposure to DI reform)	Disability insurance reform	Hospitalizations and mortality
González, L. (2013)	Spain	Date of policy change	Universal child benefit	Incidence of conceptions and abortions
Guertzgen, N. and K. Hank (2014)	Germany	Month (of child's birth relative to reform)	Reform of maternity leave legislation	Long-term sickness
Johansson, P. and M. Palme (2005)	Sweden	Date of reform	National sickness insurance	Incidence and duration of work absences
Lammers, M., H. Bloemen and S. Hochguertel (2013)	Netherlands	Age	Policy change in benefits requirements	Transition to disability benefits
Rieck, K. M. E. (2012)	Norway	Child's date of birth	Paid paternity leave	Parental sickness absence

Snyder, S. E. and W. N. Evans (2006)	USA	Quarter of birth	Lower income due to change in social security benefits ("Notch")	Five-year mortality

Table A1.5. Regression discontinuity applications in the investigation of health outcomes of social policies in developing countries. These studies investigate the indirect health effects of policy interventions and programmes that were not solely or primarily designed or intended to effect a change in the specified health outcome at the population level (see Table 3) or the individual level (see Tables 4 and 5).

Study	Context	Forcing variable	Intervention	Outcome(s)
Alam, A. and J. E. Baez (2011)	Pakistan	District literacy rate (program eligibility criterion)	Female School Stipend Program (conditional cash transfer)	sexual and fertility decisions (early marriage and childbearing)
Andalõn, M. (2011)	Mexico	Poverty index (programme eligibility criterion)	Oportunidades conditional cash transfer	rates of overweight and obesity
Bor, J. (2013)	South Africa	Date of birth	Extension of eligibility for Child Support Grant	Time to first pregnancy from age 14 (teenage pregnancy)
Carneiro, P., E. Galasso and R. Ginja (2014)	Chile	Poverty index (programme eligibility criterion)	Chile Solidario anti-poverty programme	Water and sewage connection
Carranza Barona and Mendez Sayago, 2015	Ecuador	Selben welfare index	Bono de Desarrollo Humano (conditional cash transfer)	Exclusive breastfeeding in first six months of life
Chen, Y., A. Ebenstein, M.	China	latitude relative to Huai River boundary	coal for winter heating	mortality and life expectancy

Study	Context	Forcing variable	Intervention	Outcome(s)
Greenstone and H. Li (2013)				
Cogneau, D., S. Mesple-Somps and G. Spielvogel (2013)	Cote d'Ivoire, Mali, Ghana, Guinea	distance from border	national boundaries	children's height-for-age, access to safe water
Crost, B., J. Felter and P. Johnston (2014)	Philippines	Distance of municipal poverty ranking from programme eligibility threshold	KALAHI-CIDSS, community-driven development programme	Number of conflict casualties
de Brauw, A. and A. Peterman (2011)	El Salvador	Municipal poverty score	Comunidades Solidarias Rurales (CCT)	prenatal and postnatal care, skilled attendance, birth at health facility
Filmer, D. and N. Schady (2014)	Cambodia	Dropout risk score (programme eligibility criterion)	Scholarships for poor children	Teenage pregnancy
Gordon, D. and D. L. Miller. (2012)	South Africa	Age	Old age pension eligibility	Mortality, self-reported health, access to clean water, nutrition
Janssens, W. (2011)	India	Age	Mahila Samakhya, women's empowerment and health education program	Child vaccinations
Lamadrid-Figueroa et al. (2008)	Mexico	Poverty score	Oportunidades social programme	Contraceptive use

Study	Context	Forcing variable	Intervention	Outcome(s)
Medina, C., J. Nunez and J. A. Tamayo (2013)	Colombia	Welfare index (SISBEN)	Unemployment Subsidy and retraining	Children's weight, height, BMI, Apgar score
Nabernegg 2012	Ecuador	Selben welfare index	Bono de Desarrollo Humano (conditional cash transfer)	Household spending on alcohol and cigarettes
Pitt, M.M., Khandker, S.R., McKernan, S. & Latif, M.A. (1999)	Bangladesh	Acres of land owned by household (programme eligibility criterion)	Group-based credit programmes for the poor	Contraceptive use and fertility
Rahman, M. M. (2014)	Bangladesh	Household income	Social safety net programmes	Daily caloric consumption
Siaplay, M. (2012)	South Africa	Age	South African Old Age Pension programme	Sexual behaviours of young adults in household
Sun, A. and Y. Zhao (2014)	China	Month and year of conception	Increased women's bargaining power following divorce reform	Sex ratio of second children following firstborn girls; birth spacing; child caloric intake; husband's alcohol and cigarette consumption
Tibone, K. L. (2013)	Ethiopia	Month and year of conception	US foreign aid policy change ('Mexico City Policy')	abortion rates

Study	Context	Forcing variable	Intervention	Outcome(s)
Urquieta, J., G. Angeles and T. Mroz (2009)	Mexico	Poverty index (programme eligibility criterion)	Oportunidades poverty alleviation programme	Skilled attendance at delivery
You, J. (2013)	China	Predicted probability of borrowing microcredit	Formal microcredit (Rural Credit Cooperatives)	Child malnutrition (BMI, anaemia, zinc deficiency, parent-reported health status)

Table A1.6. Regression discontinuity applications in the evaluation of health insurance schemes in developed countries.

Study	Context	Forcing variable	Exposure	Outcome(s)
Palangkaraya, A. and J. Yong (2007)	Australia	Age	Lifetime Health Cover scheme	Private health insurance coverage
Guthmuller, S. and J. Wittwer (2012)	France	Income (insurance eligibility threshold)	Universal complementary health insurance (CMU-C)	Number and probability of visits to GP/specialist/any doctor
Hullegie, P.G.J. & Klein, T.J. 2010	Germany	Income (insurance eligibility threshold)	Private health insurance	Doctor visits, nights in hospital, self-assessed health
Nishi, A., J. Michael McWilliams, H. Noguchi, H. Hashimoto, N. Tamiya and I. Kawachi (2012)	Japan	Age	Reduced copayment for low-income elderly	Physical and mental health scales; out-of-pocket medical spending
Shigeoka, H. (2014)	Japan	Age	Elderly Health Insurance programme (Japan)	Healthcare utilisation, mortality, self-reported health
Ai, E. C. Norton and Yang (2011)	USA	Age (eligibility for Medicare)	health insurance	Hospital admissions and costs
Anderson, M. L., C. Dobkin and T. Gross (2014)	USA	Age (loss of parental insurance coverage at age 23)	health insurance	ED visits, inpatient admissions
Anderson, M. L., C. Dobkin and T. Gross (2012)	USA	Age (loss of parental insurance coverage at age 19)	health insurance	ED visits, inpatient admissions
Belenkiy, M. (2010)	USA	Age (loss of parental insurance coverage at 19)	Health insurance	Obstetric treatment intensity

Study	Context	Forcing variable	Exposure	Outcome(s)
Beuermann, D. W. (2010)	USA	Age (eligibilility for Medicare)	health insurance	Healthcare utilisation/access/service quality measures
Burns, M. E., L. Dague, T. Deleire, M. Dorsch, D. Friedsam, L. J. Leininger, G. Palmucci, J. Schmelzer and K. Voskuil (2014)	USA	Date (that programme enrollment suddenly closed)	health insurance (evaluation of Medicaid expansion in Wisconsin)	ED visits, hospitalisations, outpatient visits
Card, D., C. Dobkin and N. Maestas (2009)	USA	Age (eligibilility for Medicare)	Medicare health insurance coverage	Mortality; treatment intensity
Card, D., C. Dobkin and N. Maestas (2008)	USA	Age (eligibilility for Medicare)	Medicare health insurance coverage	healthcare utilisation (multiple measures)
Card, D. and L. D. Shore-Sheppard (2004)	USA	Age	Medicaid programme expansion	Health insurance coverage
Cardella, E. and B. Depew (2014)	USA	Age	Health insurance	Self-reported health
Chay, Kim, Shailender (2010)	USA	Age	Medicare	Hospital utilisation, restricted activity, mortality
Dague, L. (2014)	USA	Family income as % of Federal Poverty Level	Medicaid/CHIP	Length of continuous enrollment

Study	Context	Forcing variable	Exposure	Outcome(s)
De La Mata, D. (2012)	USA	Family income as % of Federal Poverty Level	Medicaid	uptake, crowdout, healthcare utilisation, health status, obesity, school sickness absence
Decker, S. L. (2005)	USA	Age	Medicare eligibility	Access to mammography, stage of diagnosis, survival of breast cancer
Dugan, J., S. S. Virani and V. Ho (2012)	USA	Age (65, eligibility for Medicare)	Medicare	Physician visits, access to care, supplementary insurance coverage
Hu, T., S. L. Decker and S.-Y. Chou (2014)	USA	Age	Medicare Part D (introduction of drug coverage)	Quantity and type of drugs prescribed
Koch, T. G. (2013)	USA	Family income as a fraction of poverty guideline	public health insurance for children (SCHIP)	crowdout, healthcare utilization and spending
Muhlestein, D. B. and E. E. Seiber (2013)	USA	Family income as percentage of Federal Poverty Level	Medicaid eligibility	Crowdout of private insurance
Nikolova, S. and S. Stearns (2014)	USA	Family income as percentage of Federal Poverty Level	CHIP premium structure	insurance status
Witman, A. (2015)	USA	Age	Spousal Medicare eligibility	Insurance coverage of younger spouse (crowd-out)

Table A1.7. Regression discontinuity applications in the evaluation of health insurance schemes in developing countries.

Study	Context	Forcing variable	Exposure	Outcome(s)
Camacho, A. and E. Conover (2013)	Colombia	Poverty index (programme eligibility criterion)	Subsidized Regime (SR) health insurance for the poor	Newborn health (LBW, VLBW, Apgar 5), prenatal care
Miller, G., D. Pinto and M. Vera-Hernández (2013)	Colombia	Simulated SISBEN index	Subsidised Regime of health insurance for the poor	Service use, health status, health behaviours
Bauhoff, S., D. R. Hotchkiss and O. Smith (2011)	Georgia	Programme eligibility score (based on >100 household indicators)	Medical Insurance Program for the Poor (MIP)	Healthcare utilisation, out of pocket expenditure, individual health status and behaviours
Hou, X. and S. Chao (2008)	Georgia	Welfare score	Medical Assistance Program for the poor	Acute surgeries and inpatient care
Sood, N., E. Bendavid, A. Mukherji, Z. Wagner, S. Nagpal and P. Mullen (2014)	India	Geographic boundary	Public health insurance (tertiary care for households below poverty line)	Mortality, healthcare utilization, out-of-pocket expenditure
Bernal, N., M. A. Carpio and T. J. Klein (2014)	Peru	Welfare index (programme eligibility threshold)	Peruvian social health insurance	Healthcare utilisation, expenditure, individual health outcomes
Yang, T.-T., H.-W. Han and H.-M. Lien (2014)	Taiwan	Age	Taiwan Children's Medical Subsidy Program	Healthcare utilization and expenditure

Palmer, M., S. Mitra, D. Mont and N. Groce (2014)	Vietnam	Age	Public health insurance for preschool children	Inpatient and outpatient visits (healthcare utilisation), expenditure, substitution(crowdout)

Table A1.8. Regression discontinuity applications in epidemiological questions of cause and effect.
These studies investigate the health effects of exposures that were not part of a social, clinical, or public health intervention, programme, or policy.

Study	Context	Forcing variable	Exposure	Outcome(s)
Bhalotra, S., I. Clots-Figueras, G. Cassan and L. Iyer (2014)	India	Vote margin in close elections	Rise in share of elected officials who are Muslim	Neonatal and infant mortality
Conley, D. and J. Heerwig (2012)	USA	Lottery number cutoff for draft eligibility	Vietnam War military conscription	Mortality
Cullen, K. W., L. M. Koehly, C. Anderson, et al. (1999)	USA	years from age 18	Transition from high school	diet, physical activity, tobacco and alcohol use, sexual behaviour
Dell, M. (2010)	Peru	latitude and longitude	the mita, a forced labour system in operation 1573-1812	stunted growth in children
Dickert-Conlin, S. and T. Elder (2010)	USA	Date (state cutoff for school eligibility)	Cutoff dates for starting school	Share of annual births by calendar day
Eibich, P. (2014)	Germany	Age	Retirement	Physical and mental health, smoking, alcohol, exercise, diet, sleep, social support, healthcare utilization
Fé, E. and B. Hollingsworth (2012)	UK	Default retirement age	Retirement	Mental health indicators, healthcare

Study	Context	Forcing variable	Exposure	Outcome(s)
				utilisation, BP, migraine
Fletcher, J. M. (2014)	USA	Date of survey interview	September 11th terror attacks	Sadness
Huang, W. and Y. Zhou (2013)	China	Born in 1948	Great Famine 1959-61	Cognitive functioning
Johnston, D. W. and W. S. Lee (2009)	UK	Age	Retirement	GHQ-12 mental health, BMI, hypertension
Kong, A. (2011)	Canada	Age	Retirement	Self-reported physical and mental health
Pierce, L., M. S. Dahl and J. Nielsen (2013)	Denmark	Marital income difference	Income inequality between spouses	Prescription medications for erectile dysfunction, anxiety, insomnia, depression
Sotomayor, O. (2013)	Puerto Rico	Year of birth	In-utero exposure to natural disasters (hurricanes)	Hypertension, diabetes, high cholesterol in adulthood
Zhong, H. (2014)	China	Year of birth	Number of siblings	Child health (height, self-assessed health, BMI)

Table A1.9. Regression discontinuity applications in the causal impact of education on health.
These studies investigate the health effects of education programmes, exposure to education, and changes in educational policy.

Study	Context	Forcing variable	Exposure	Outcome(s)
Albouy, V. and L. Lequien (2009)	France	Year of policy change	Raised mandatory minimum school leaving age	Mortality (survival rates at age 50 and 80)
Anderson, P. M., K. F. Butcher, E. U. Cascio and D. W. Schanzenbach (2011)	USA	Birthdate (cutoff for starting school)	Years of early primary education	BMI
Arcand, J. L. and E. D. Wouabe (2010)	Cameroon	Number of secondary schools in town (programme eligibility criterion)	HIV/AIDS teacher training programme	HIV-related knowledge, attitudes and behaviour
Banks, J. and F. Mazzonna (2012)	UK	Birthdate	1947 policy change in minimum school leaving age (additional year of schooling)	Memory, executive functioning, CASP-19, social and cultural activity index
Behrman, J. A. (2015)	Malawi and Uganda	Birth cohort	Universal Primary Education	HIV status
Clark, D. and H. Royer (2013)	UK	birthdate (month and year)	Changes to UK compulsory schooling laws	mortality, health behaviours, self-reported health
Greenwood, E. (2012)	USA	Year	College opening	Births to teenage mothers
Jakobsson, N., M. Persson and M. Svensson (2013)	Sweden	Class size	Class size	Mental health and wellbeing measures

Study	Context	Forcing variable	Exposure	Outcome(s)
Johnston, D.	UK	Date of birth	Additional year of schooling	Index of health knowledge
Jurges, H., E. Kruk and S. Reinhold (2010)	UK	Date of birth	Additional year of schooling	Blood fibrinogen, CRP, self-reported health
Lindeboom, M., A. Llena-Nozal and B. van der Klaauw (2009)	UK	Year of birth	Additional year of schooling	Child height, weight, morbidity; parental BMI, chronic disease, fertility
Lleras-Muney, A. (2005)	USA	Year of change in compulsory schooling education	Education	Mortality
McCrary, J. and H. Royer (2011)	USA	Date of birth	School starting age	Fertility, birthweight and prematurity
Monstad, K., C. Propper and K. G. Salvanes (2008)	Norway	Age relative to year of reform	Reform that increased years of compulsory schooling	Number of children and maternal age at first birth
Nakamura, R. (2012)	UK	Month and year of birth	Maternal schooling	Children's bodyweight, fruit and veg consumption, exercise
Park, W. (2013)	South Korea	Year of birth	College education	Smoking behaviour
Powdthavee, N. (2010)	UK	Year of birth	Compulsory education	Hypertension
Samarakoon, S. and R. A. Parinduri (2015)	Indonesia	Year of birth	Education (longer school year in 1978)	Fertility and reproductive health behaviours

Study	Context	Forcing variable	Exposure	Outcome(s)
Silles, M.A. (2009)	UK	Unclear (age or year)	Years of schooling	Self-reported health
van Kippersluis, H., O. O'Donnell and E. van Doorslaer (2011)	Netherlands	Birthdate	Years of compulsory schooling	Mortality after age 81
Zhang, N. (2009)	USA	Age	Years of formal schooling	Children's bodyweight, fruit and vegetable consumption
Zhong, H. (2015)	China	Date	College education	Smoking, drinking, self-rated health, hypertension, weight

Appendix 3. Detailed critical appraisal results for chapter 4

This appendix reports the detailed results for quality assessment of the 17 regression discontinuity studies of minimum legal drinking age legislation included in chapter 4 using the What Works Clearinghouse Standards for RD. Each study was appraised independently by two reviewers. The results shown below are the final consensus assessments agreed upon following discussion.

The study ID is the first four letters of the first author's name plus the year of publication and first page number (or WP for working paper). WWC 1, 2, and 3 are the qualifying questions. Columns labelled with numbers (1a, 1b, etc) refer to criteria and columns labelled with S refer to standards. Assessments were coded '1' to mean the criteria or standard was met, '0' to mean it was not met, and 'MWR' to mean 'met with reservations'. 'NA' means not applicable.

Study ID	WWC1	WWC2	WWC3	1a	1b	S2	3a	3b
BOES_2013_WP	1	1	1	1	0	0	NA	1
CALL_2014_788 a	1	1	1	1	0	0	NA	1
CALL_2013_1590 a	1	1	1	1	0	0	NA	1
CALL_2013_2284 b	1	1	1	1	0	0	NA	1
CALL_2014_137 b	1	1	1	1	0	0	NA	0
CARP_2009_164	1	1	1	1	0	0	1	0
CARP_2011_133	1	1	1	1	0	0	NA	0
CARP_2015_WP	1	1	1	1	0	0	NA	0
CONO_2013_570	1	1	1	1	0	0	NA	1
CROS_2012_112	1	1	1	1	0	0	NA	1

Study ID	WWC1	WWC2	WWC3	1a	1b	S2	3a	3b
CROS_2013_474	1	1	1	1	0	0	NA	0
DEZA_2015_419	1	1	1	1	1	1	1	1
ERTA_2015_133	1	1	1	1	0	0	1	0
ERTA_2012_1844	1	1	1	1	0	0	1	0
LIND_2014_WP	1	1	1	1	1	0	1	0
YORU_2011_740	1	1	1	1	1	0	1	0
CARP_2014_WP	1	1	1	1	1	0	1	0

Study ID	4a	4b	4c	4e	S1	S3	S4	Overall	Notes
BOES_2013_WP	1	1	1	NA	1	1	1	1	
CALL_2014_788 a	1	1	1	NA	1	1	1	1	
CALL_2013_1590 a	1	1	1	0	1	1	MWR	MWR	Results not presented by site (Province) when could have been - 4e
CALL_2013_2284 b	1	1	1	NA	1	1	1	1	

Study ID	4a	4b	4c	4e	S1	S3	S4	Overall	Notes
CALL_2014_137 b	1	1	1	1	1	0	1	MWR	Because 3b not met
CARP_2009_164	1	1	1	NA	1	0	1	MWR	Because 3b not met
CARP_2011_133	1	1	1	NA	1	0	1	MWR	Because 3b not met
CARP_2015_WP	1	1	1	NA	1	0	1	MWR	Because 3b not met
CONO_2013_570	1	1	1	NA	1	1	1	1	
CROS_2012_112	1	1	1	NA	1	1	1	1	
CROS_2013_474	1	1	1	NA	1	0	1	MWR	Because 3b not met
DEZA_2015_419	1	1	1	NA	1	1	1	1	
ERTA_2015_133	1	1	1	NA	1	0	1	MWR	Because 3b not met
ERTA_2012_1844	1	1	1	NA	1	0	1	MWR	Because 3b not met

Study ID	4a	4b	4c	4e	S1	S3	S4	Overall	Notes
LIND_2014_WP	1	0	1	NA	1	0	MWR	MWR	Because 3b not met and 4b graphs don't have fitted curves
YORU_2011_740	1	1	1	NA	1	0	1	MWR	Because 3b not met
CARP_2014_WP	1	1	1	NA	1	0	1	MWR	Because 3b not met

Appendix 4. Protocol: Endocrine disrupting chemicals and the risk of breast cancer: a systematic review of reviews [CRD42018089344]

Endocrine disrupting chemicals and the risk of breast cancer: a systematic review of reviews

Michele Hilton Boon, Laurence Moore, Hilary Thomson, Peter Craig

Citation

Michele Hilton Boon, Laurence Moore, Hilary Thomson, Peter Craig. Endocrine disrupting chemicals and the risk of breast cancer: a systematic review of reviews. PROSPERO 2018 CRD42018089344 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018089344

Review question

1. What is the evidence from systematic reviews that endocrine disrupting compounds (EDCs) increase the risk of breast cancer in humans?

2. What is the contribution of natural experiments to the evidence base on the causal role of EDCs in breast cancer?

3. How have systematic reviews evaluated and presented evidence from different study designs, including natural experiments, in reaching their conclusions about EDCs?

4. How do systematic reviews of EDCs and breast cancer vary in their methodology with respect to inclusion criteria, appraisal methods, and synthesis methods, and how do these variations affect the inclusion and presentation of results from natural experiments?

5. What have systematic reviews identified as limitations and gaps relating to natural experiments within the evidence base on EDCs and breast cancer in humans?

Searches

The databases to be searched are MEDLINE, Embase, the Cochrane Database of Systematic Reviews (CDSR), BIOSIS Previews, Scopus, and Web of Science.

Additionally, Google and OpenGrey will be searched for relevant grey literature, and IARC monographs will be searched (http://monographs.iarc.fr/).

The search strategy for the bibliographic databases will combine terms for endocrine disruptors and breast cancer with a filter to identify systematic reviews.

This meta-review will include systematic reviews published on or after 1st January 2003, the search cut-off dates for which are no earlier than 1st January 2002. The year 2002 has been chosen because it was the date of the publication

of the first Global Assessment of the State of the Science of Endocrine Disruptors (International Programme on Chemical Safety, 2002).

No language restrictions will be imposed at the search stage.

Types of study to be included

This meta-review will include systematic reviews, defined as a review that (1) follows a specific, transparently reported, reproducible method of retrieving and selecting studies in an effort to comprehensively address its research question, and (2) presents the characteristics and results of included papers in some form of synthesis (quantitative, qualitative, or narrative).

'Empty' reviews (reviews that identified no studies that met the inclusion criteria) will be included, but the protocols of reviews that have not reported any findings will be excluded. Primary studies will not be included either.

In addition, included reviews must have addressed (at least in part, but not necessarily exclusively) the PICO question of the effect in humans (P) of exposure to EDCs (I) compared with any variation in exposure, degree, or timing (C) on the risk of breast cancer (O).

Condition or domain being studied

Breast cancer and its environmental causes.

Participants/population

Humans exposed to endocrine disrupting chemicals.

Intervention(s), exposure(s)

The exposure of interest is endocrine disrupting chemicals.

The WHO/IPCS definition states "An endocrine disrupter is an exogenous substance or mixture that alters function(s) of the endocrine system and consequently causes adverse health effects in an intact organism, or its progeny, or (sub)populations."

Known EDCs include dioxins, polychlorinated biphenyls (PCBs), certain pesticides, herbicides, fungicides, and consumer product chemicals such as bisphenol A, phthalates, nonylphenols, flame retardants, and organic solvents.

Environmental, household, and occupational exposures are also included, and alcohol and benzene are included in the category of organic solvents, but the common routes of exposure to these (alcohol consumption, and benzene in tobacco smoke) will be excluded. Pharmaceuticals will also be excluded.

Comparator(s)/control

The comparators may be any variation in exposure (including non-exposure), degree, or timing.

Context

Primary outcome(s)

The primary quantitative outcome of the review is the risk of breast cancer in humans, expressed as relative risk (RR), odds ratio (OR), or hazard ratio (HR), associated with a given exposure to an EDC or combination of EDC under a given set of circumstances, with 95% confidence intervals.

The primary qualitative outcome of the review is a map of evidence that demonstrates (1) the number and type of natural experimental studies included in the evidence base and (2) the amount of overlap of included studies among the systematic reviews.

Secondary outcome(s)

Sensitivity analyses of the primary outcomes by review characteristics, review quality (AMSTAR2 score), and the inclusion of natural experiments.

Data extraction (selection and coding)

A data extraction form will be designed and piloted on two systematic reviews (one reviewer will pilot the data extraction form and a second will cross-check the extracted data for accuracy). The data extraction form will be revised if necessary and revisions will be reported with explanations for any changes.

One reviewer will then extract relevant data from all eligible studies, and a second will cross-check the extracted data for accuracy. Any disagreements will be resolved through discussion or, if necessary, with the involvement of a third reviewer.

The data to be extracted from each included review will be:

Review characteristics: the citation, year of publication, objectives, search cut-off date, databases searched, inclusion criteria, quality appraisal method, method(s) of synthesis.

Details of the included studies: number of studies and population numbers included in the review, references of included studies (human populations only), number and date range of other included studies (animal and in vitro), designs of included studies in humans.

Details of the review findings: EDCs covered, characteristics of EDC exposure covered (doses, timeframes, modifying factors), results of meta-analysis of risk of cancer in humans, numeric estimates of risk from included natural experiments in human populations, results of narrative synthesis, overall assessment of risk of bias and/or certainty of evidence, limitations or gaps noted in the evidence base.

Data will not be extracted from the primary studies included in the reviews. In case of any discrepancies between the reviews, (e.g., different reports of study characteristics or results for a study included in multiple reviews), all data will be recorded but discrepancies will be highlighted and erroneous data excluded from further synthesis. If there is found to be data missing from the included primary studies, the data will be reported as missing, and the data left incomplete.

Risk of bias (quality) assessment

Included systematic reviews will be critically appraised using the AMSTAR2 checklist, which has been developed for the appraisal of systematic reviews that may include evidence from both randomised and non-randomised studies.

Two reviewers will appraise each study independently, and any disagreements will be resolved through discussion. The appraisal results will be presented in a table and in a summary chart.

The AMSTAR2 checklist is intended for the appraisal of systematic reviews of intervention studies in healthcare, so in order to ensure that appropriate consideration is given to criteria specific to epidemiological studies which examine the effects of exposures rather than interventions, reviews that include a meta-analysis of risk of breast cancer in humans will additionally be appraised using the MOOSE (Meta-analyses Of Observational Studies in Epidemiology) checklist.

Strategy for data synthesis

The focus of the synthesis will be on findings from natural experiments in human populations because of the potential of these study designs to contribute to understanding causality.

Included systematic reviews are likely to contain a wide range of types of evidence, including in vitro data, experiments on animals, and observational studies on wildlife. If decisions need to be made about the level of detail or depth of the synthesis, the emphasis will be on representing findings from studies of human populations. We will describe in tables the characteristics of the included reviews and the exposures of interest that they address.

The primary quantitative outcome (risk of breast cancer) will be presented as a forest plot of review results ordered by date, with estimates of risk associated with different exposures presented separately where possible.

The primary qualitative outcome (map of evidence) will be presented graphically and narratively.

Overlaps of primary studies among reviews will be presented in a tabular format and described narratively.

A thematic analysis will be used to further investigate the map of evidence in order to identify the contribution of natural experiments, limitations and gaps in the evidence base, and points of comparison with the content of policy documents. These results will be presented graphically, if feasible, and narratively.

Analysis of subgroups or subsets

We will separately analyse different groupings of EDCs depending on the review coverage (for example, as persistent versus non-persistent EDCs, or groupings such as dioxins, organochlorine pesticides, phthalates).

We will separately present, explore, and analyse the findings from natural experiments.

Contact details for further information

Michele Hilton Boon

m.boon@sphsu.mrc.ac.uk

Organisational affiliation of the review

MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccsosocialandpublichealthsciencesunit/

Review team members and their organisational affiliations

Ms Michele Hilton Boon. University of Glasgow
Professor Laurence Moore. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Dr Hilary Thomson. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow
Dr Peter Craig. MRC/CSO Social and Public Health Sciences Unit, University of Glasgow

Anticipated or actual start date

01 February 2018

Anticipated completion date

30 June 2018

Funding sources/sponsors

LM is supported by the UK Medical Research Council (MC_UU_12017/14) and the Scottish Government Chief Scientist Office (SPHSU14). HT and PC are core funded by the UK Medical Research Council (MC_UU_12017/13 & MC_UU_12017/15) and the Scottish Government Chief Scientist Office (SPHSU13 & SPHSU15). MHB is funded by a UK Medical Research Council doctoral studentship (Natural experimental approaches to evaluating population health interventions: 1517742)

Conflicts of interest

None specified.

Language

(there is not an English language summary)

Country

Scotland

Stage of review

Review_Ongoing

Subject index terms status

Subject indexing assigned by CRD

Subject index terms

Breast Neoplasms; Endocrine Disruptors; Environmental Exposure; Environmental Pollutants; Environmental Pollution; Humans; Risk; Risk Factors

Date of registration in PROSPERO

27 February 2018

Date of publication of this version

27 February 2018

Appendix 5. Literature search strategies for chapter 6

Search Concepts:

breast cancer + systematic review + endocrine disrupting chemicals

Sources of Terms:

Based on Rodgers et al. (2018) search strategy following Brody et al. (2007) and Rudel et al. (2014) with additional synonyms drawn from the following sources:

Rachoń D. Endocrine disrupting chemicals (EDCs) and female cancer: Informing the patients. *Reviews in Endocrine & Metabolic Disorders*. 2015;16:359-364.

Gore AC, Chappell VA, Fenton SE, et al. EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. *Endocrine Reviews*. 2015;36(6):E1-E150. doi:10.1210/er.2015-1010.

EU report on EDC identification and categorisation:
ec.europa.eu/environment/chemicals/endocrine/strategy/substances_en.htm
Terms from Annex 13, The Summary Profiles of (41) Category 1 Chemical Groups

Search Strategies:

Medline (18 March 2018, Ovid platform)

1. exp Breast Neoplasms/
2. (breast$ or mammary).mp.
3. (cancer$ or tumo?r or neoplasm$).mp.
4. 2 and 3
5. 1 or 4
6. exp Endocrine Disruptors/
7. (endocrine adj disrupt$).mp.
8. *Phenols/
9. bisphenol A.mp.
10. ddt/ or dichlorodiphenyl dichloroethylene/ or dichlorodiphenyldichloroethane/
11. exp "DIOXINS AND DIOXIN-LIKE COMPOUNDS"/
12. dioxin$.mp.
13. Flame Retardants/
14. flame retardant$.mp.
15. exp Fungicides, Industrial/
16. fungicid$.mp.
17. exp HERBICIDES/
18. herbicid$.mp.
19. exp INSECTICIDES/
20. insecticid$.mp.
21. paraben$.mp.
22. exp PARABENS/
23. exp Paraffin/
24. exp Polychlorinated Biphenyls/
25. PCBs.mp.

26. exp PESTICIDES/
27. pesticid$.mp.
28. DIETHYLHEXYL PHTHALATE/ or DIBUTYL PHTHALATE/
29. phthalate$.mp.
30. exp Surface-Active Agents/
31. surfactant$.mp.
32. or/6-31
33. 5 and 32
34. limit 33 to (meta analysis or systematic reviews)
35. limit 33 to "reviews (best balance of sensitivity and specificity)"
36. 34 or 35
37. Aldrin.mp. or ALDRIN/
38. alkylphenol.mp.
39. Araclor.mp.
40. Atrazine.mp. or ATRAZINE/
41. BADGE.mp.
42. BBMP.mp.
43. exp BENZENE DERIVATIVES/ or exp BENZENE/ or benzene.mp.
44. benzophenone-1.mp.
45. exp Pyrethrins/ or bifenthrin.mp.
46. BPA.mp.
47. Captan/ or Captafol.mp.
48. Carbaryl.mp. or CARBARYL/
49. carbamate.mp. or exp CARBAMATES/
50. chlordane.mp. or Chlordan/
51. Chlordecone.mp. or CHLORDECONE/
52. chloroparaffin$.mp.
53. chlorotriazine.mp.
54. chlorpyrifos.mp. or CHLORPYRIFOS/
55. cyhalothrin.mp.
56. DDD.mp.
57. DDE.mp.
58. decaBDE$.mp.
59. DEHP.mp. or Diethylhexyl Phthalate/
60. deltamethrin.mp.
61. DEP.mp.
62. exp Detergents/
63. detergent$.mp.
64. diazinon.mp. or DIAZINON/
65. dicarboximide.mp.
66. dichlorodiphenyldichloroethane.mp. or DICHLORODIPHENYLDICHLOROETHANE/
67. Dichlorodiphenyl Dichloroethylene/ or dichlorodiphenyldichloroethylene.mp.
68. dichlorophenyldichloroethylene.mp.
69. dicofol.mp. or DICOFOL/
70. exp Phthalic Acids/ or dicyclohexylphthalate.mp.
71. diethylphthalate.mp.
72. Dieldrin.mp. or DIELDRIN/
73. DnBP.mp.
74. Ethylene Dibromide/ or EDB.mp.
75. Endrin.mp. or ENDRIN/
76. epichlorohydrin.mp. or EPICHLOROHYDRIN/

77. Ethanol.mp. or ETHANOL/
78. Fenarimol.mp.
79. exp Pyrimidines/
80. fenitrothion.mp. or FENITROTHION/
81. fenvalerate.mp.
82. fluorosurfactant$.mp.
83. Hair dye$.mp. or exp Hair Dyes/
84. exp Hair Preparations/ or Hair relaxer$.mp.
85. exp Cosmetics/ or Hair straightener$.mp.
86. HCB.mp. or Hexachlorobenzene/
87. HCH$.mp.
88. HEPTACHLOR EPOXIDE/ or HEPTACHLOR/ or heptachlor.mp.
89. hexachlorobenzene.mp. or HEXACHLOROBENZENE/
90. Hexachlorohexane.mp.
91. ioxynil.mp.
92. Kanechlor.mp.
93. lindane.mp. or LINDANE/
94. malathion.mp. or MALATHION/
95. mancozeb.mp.
96. Diethylhexyl Phthalate/ or MEHP.mp.
97. methoxychlor.mp. or METHOXYCHLOR/
98. methylene chloride.mp. or Methylene Chloride/
99. metiram.mp.
100. metribuzin.mp.
101. Mirex.mp. or MIREX/
102. exp Hydrocarbons, Chlorinated/ or nonachlor.mp.
103. nonylphenol$.mp.
104. exp Phenols/
105. exp Halogenated Diphenyl Ethers/ or octaBDE.mp.
106. ?octylphenol$.mp.
107. exp Solvents/
108. Organochlorine.mp.
109. exp Polycyclic Aromatic Hydrocarbons/
110. polycyclic aromatic hydrocarbon$.mp.
111. PARATHION/ or METHYL PARATHION/ or parathion.mp.
112. PBDE$.mp.
113. PCB$.mp.
114. pentaBDE$.mp. or exp Hydrocarbons, Brominated/
115. pentachlorobenzene.mp.
116. pentachlorophenol.mp.
117. Perfluoroalkyl$.mp.
118. perfluorooctanesulfonic acid.mp.
119. perfluorooctanoic acid.mp.
120. *Environmental Pollutants/ or persistent organic pollutant$.mp. or *Water Pollutants, Chemical/
121. PFASs.mp.
122. (PFOA or PFOS or PHDD$ or PHDF$).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
123. picloram.mp. or PICLORAM/
124. plastici?er.mp. or *Plasticizers/

125. (polychlorinated or polybrominated or polyfluorinated).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
126. Polyfluoroalkyl$.mp.
127. polyvinyl chloride.mp. or Polyvinyl Chloride/
128. procymidone.mp.
129. PVC.mp.
130. exp Pyrethrins/
131. (pyrethroid$ or pyretroid$).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
132. resmethrin.mp.
133. simazine.mp. or SIMAZINE/
134. (TCDD or TCE or TDBPP).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
135. Terbufos.mp.
136. terbutryn.mp.
137. tetraBDE47.mp.
138. tetrachloroethylene.mp. or TETRACHLOROETHYLENE/
139. Toxaphene.mp. or TOXAPHENE/
140. exp *Chlorobenzenes/ or trichlorobenzene.mp.
141. trichloroethylene.mp. or TRICHLOROETHYLENE/
142. vinclozolin.mp.
143. or/37-142
144. 5 and 143
145. limit 144 to (meta analysis or systematic reviews)
146. limit 144 to "reviews (best balance of sensitivity and specificity)"
147. 36 or 145 or 146
148. limit 147 to yr="2003 -Current"
149. 37 or 38 or 39 or 40 or 41 or 42 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75 or 76 or 78 or 80 or 81 or 82 or 83 or 84 or 85 or 86 or 87 or 88 or 89 or 90 or 91 or 92 or 93 or 94 or 95 or 96 or 97 or 98 or 99 or 100 or 101 or 102 or 103 or 104 or 105 or 106 or 107 or 108 or 110 or 111 or 112 or 113 or 114 or 115 or 116 or 117 or 118 or 119 or 120 or 121 or 122 or 123 or 124 or 125 or 126 or 127 or 128 or 129 or 130 or 131 or 132 or 133 or 134 or 135 or 136 or 137 or 138 or 139 or 140 or 141 or 142
150. 5 and 149
151. limit 150 to (meta analysis or systematic reviews)
152. limit 150 to "reviews (best balance of sensitivity and specificity)"
153. 36 or 151 or 152
154. limit 153 to yr="2003 -Current"
Sources: Ovid MEDLINE(R) without Revisions 1996 to March Week 2 2018
Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) without Revisions 2014 to Daily Update

Embase strategy (18 March 2018, Ovid platform)

File: Embase 1996 to 2018 Week 12

1. exp breast cancer/ or breast tumor/
2. (breast$ or mammary).mp.
3. (cancer$ or tumo?r or neoplasm$).mp.
4. 2 and 3
5. 1 or 4
6. exp endocrine disruptor/
7. (endocrine adj disrupt$).mp.
8. 4,4' isopropylidenediphenol/
9. bisphenol A.mp.
10. DDT.mp. or chlorphenotane/
11. dioxin/
12. dioxin$.mp.
13. flame retardant/
14. flame retardant$.mp.
15. fire retardant$.mp.
16. exp fungicide/
17. fungicid$.mp.
18. exp herbicide/
19. herbicid$.mp.
20. exp insecticide/
21. insecticid$.mp.
22. paraben$.mp.
23. 4 hydroxybenzoic acid ester/
24. paraffin/
25. polychlorinated biphenyl/
26. PCB$.mp.
27. exp polychlorinated dibenzodioxin/
28. exp pesticide/
29. pesticid$.mp.
30. "phthalic acid bis(2 ethylhexyl) ester"/ or exp plasticizer/
31. surfactant/ae, it, to, ec [Adverse Drug Reaction, Drug Interaction, Drug Toxicity, Endogenous Compound]
32. surfactant$.mp.
33. or/6-32
34. 5 and 33
35. MEDLINE.tw.
36. exp systematic review/
37. systematic review.tw.
38. meta-analysis/
39. limit 34 to (meta analysis or "systematic review")
40. 35 or 36 or 37 or 38
41. 34 and 40
42. 39 or 41
43. alkylphenol.mp.
44. BADGE.mp.
45. BBMP.mp.
46. benzophenone derivative/ or exp benzene derivative/
47. exp pyrethroid/

48. detergent/
49. dimpylate/ or exp pyrimidine derivative/
50. exp phthalic acid derivative/
51. exp 1,2 dibromoethane/
52. exp alkane derivative/
53. epichlorohydrin/ or exp epoxide/ or exp organochlorine derivative/
54. hair dye/
55. Hair dye$.mp.
56. exp cosmetic/
57. exp phenol derivative/
58. chlorinated hydrocarbon/
59. dichloromethane/ or exp organic solvent/
60. diphenyl ether derivative/
61. exp polybrominated diphenyl ether/
62. PBDE$.tw.
63. pentaBDE.mp.
64. brominated hydrocarbon/
65. pentachlorobenzene.mp. or pentachlorobenzene/
66. pentachlorophenol.mp. or pentachlorophenol/
67. perfluorooctanesulfonic acid/ or perfluorooctanoic acid/
68. *pollutant/
69. polyvinylchloride/
70. PVC.tw.
71. 2,3,7,8 tetrachlorodibenzo para dioxin/
72. dibenzodioxin derivative/ or 3,7,8 trichloro 2 iododibenzo para dioxin/ or polybrominated dibenzodioxin/ or polychlorinated dibenzodioxin/
73. (TCDD or TCE or TDBPP).tw.
74. tetrachloroethylene/
75. tetrachloroethylene.tw.
76. *persistent organic pollutant/
77. trichloroethylene.tw.
78. or/43-77
79. 5 and 78
80. 40 and 79
81. 41 or 80
82. limit 81 to yr="2003 -Current"

Cochrane Library (March 2018)

1. "breast cancer":ti,ab,kw (Word variations have been searched)
2. hormone disrupt*
3. endocrine disrupt*
4. environment*
5. chemical*
6. #4 and #5
7. #2 or #3 or #6
8. #1 and #7

Scopus (March 2018)

((TITLE-ABS-KEY ("endocrine disrupt*" OR bisphenol OR dioxin* OR pesticid* OR insecticid* OR herbicid* OR fungicid*) OR TITLE-ABS-KEY (solvent* OR plasticiser* OR plasticizer* OR surfactant* OR paraben*) OR TITLE-ABS-KEY ("DDT" OR "PCB*" OR "flame retardant*" OR "consumer product*"))) AND (((breast* OR mammary) AND (cancer* OR tumor* OR tumour* OR neoplasm*))) AND (("systematic review" OR "meta analysis" OR "meta-analysis"))

Web of Science (March 2018)

1 TS=((breast* OR mammary) AND (cancer* OR tumor* OR tumour* OR neoplasm*))

2 TS=("systematic review" OR "meta analysis" OR "meta-analysis" OR "Medline")

3 #2 AND #1

4 TS=((endocrine OR hormon*) AND (disrupt*))

5 TS=(bisphenol OR dioxin* OR pesticid* OR insecticid* OR herbicid* OR fungicid*)

6 TS=(solvent* OR plasticiser* OR plasticizer* OR surfactant* OR paraben*)

#7 TS=("DDT" OR "PCB*" OR "flame retardant*" OR "consumer product*")

#8 #7 OR #6 OR #5 OR #4

#9 #8 AND #3

Timespan=2003-2018

Open Grey (March 2018)

endocrine disrupt* breast cancer (2 results)
OR
breast cancer environment* chemical* (1 result)
OR
breast cancer environment* (26 results)

No systematic reviews identified

Google (March 2018)

endocrine disruptor breast cancer review site:.int (403 results)

endocrine disruptor breast cancer review site:.eu (1980 results; reviewed first 100 then revised search to “systematic review” which produced 319 results)

endocrine disruptor breast cancer review site:.org (1980 results; revised search to “systematic review” which produced 24,600 results; reviewed first 100)

endocrine disruptor breast cancer “systematic review” site:.gov.uk (52 results)

Appendix 6. Detailed critical appraisal results for chapter 6

This appendix reports the detailed results for quality assessment of the 15 reviews included in chapter 6 using the AMSTAR-2 appraisal tool for systematic reviews. Some of the criteria apply only to meta-analyses (11/15 included reviews); where these criteria did not apply, "N/A" has been recorded (not applicable). Each study was appraised independently by two reviewers. The results shown below are the final consensus assessments agreed upon following discussion.

Study	Rodgers 2018	Gray 2017	Brody 2007	Mouly 2016	Leng 2016
1. PICO	Yes	No	Yes	Yes	Yes
2. Protocol	No	No	No	No	No
3. Inclusion criteria	Yes	Yes	Yes	Yes	Yes
4. Search	No	Partial Yes	No	Partial Yes	Partial Yes
5. Duplicate selection	Yes	No	No	No	No
6. Duplicate extraction	Yes	No	No	Yes	Yes
7. List of excluded studies	No	No	No	No	Yes
8. Included studies described	Yes	No	Partial Yes	Yes	Yes
9. RoB assessed	Partial Yes	No	Partial Yes	Partial Yes	Yes
10. Study funding reported	No	No	No	No	No
11. Appropriate meta-analysis	N/A	N/A	N/A	N/A	Yes
12. Impact of RoB on meta-analysis	N/A	N/A	N/A	N/A	Yes
13. RoB in interpretation/discussion	Yes	No	Yes	Yes	Yes
14. Heterogeneity investigated	N/A	N/A	N/A	N/A	Yes
15. Publication bias investigated	N/A	N/A	N/A	N/A	Yes
16. COI and funding disclosed	Partial Yes	Yes	No	No	Yes
Overall confidence	Critically low	Critically low	Critically low	Critically low	Low
Highlighted domains are "critical"					

Study	Zhang 2015	Zani 2013	Allam 2016	Fu 2017	Gera 2018
1. PICO	Yes	Yes	Yes	Yes	Yes
2. Protocol	No	No	No	No	No
3. Inclusion criteria	Yes	Yes	Yes	Yes	Yes
4. Search	Partial Yes	No	Yes	Partial Yes	Partial Yes
5. Duplicate selection	Yes	No	No	No	No
6. Duplicate extraction	Yes	Yes	Yes	Yes	No
7. List of excluded studies	No	No	No	No	Partial Yes
8. Included studies described	Yes	Yes	Partial Yes	Yes	Yes
9. RoB assessed	Yes	No	No	Yes	No
10. Study funding reported	No	No	No	No	No
11. Appropriate meta-analysis	Yes	Yes	Yes	Yes	No
12. Impact of RoB on meta-analysis	Yes	No	Yes	Yes	No
13. RoB in interpretation/discussion	Yes	No	Yes	Yes	No
14. Heterogeneity investigated	Yes	Yes	No	Yes	No
15. Publication bias investigated	Yes	No	No	Yes	Yes
16. COI and funding disclosed	Yes	No	No	Yes	No
Overall confidence	Critically low	Critically low	Critically low	Critically low	Critically low
Highlighted domains are "critical"					

Study	Hardefeldt 2013	Takkouche 2005	Ingber 2013	Khanjani 2007	Park 2014
1. PICO	Yes	Yes	Yes	Yes	Yes
2. Protocol	No	No	No	No	No
3. Inclusion criteria	Yes	Yes	Yes	Yes	Yes
4. Search	Partial Yes	Partial Yes	Partial Yes	Partial Yes	Partial Yes
5. Duplicate selection	No	No	No	No	Yes
6. Duplicate extraction	No	No	Yes	Yes	Yes
7. List of excluded studies	No	Yes	Yes	Yes	No
8. Included studies described	No	Yes	Yes	Partial Yes	Partial Yes
9. RoB assessed	No	Partial Yes	No	No	No
10. Study funding reported	No	No	No	No	No
11. Appropriate meta-analysis	No	Yes	Yes	Yes	Yes
12. Impact of RoB on meta-analysis	No	Yes	No	No	No
13. RoB in interpretation/discussion	No	Yes	No	No	No
14. Heterogeneity investigated	No	Yes	Yes	Yes	Yes
15. Publication bias investigated	No	Yes	Yes	Yes	Yes
16. COI and funding disclosed	No	Yes	Yes	Yes	Yes
Overall confidence	Critically low	Low	Critically Low	Critically low	Critically low
Highlighted domains are "critical"					

References

ACADEMY OF MEDICAL SCIENCES & RUTTER, M. 2007. Identifying the environmental causes of disease: how should we decide what to believe and when to take action? London: Academy of Medical Sciences.

ACHESON, D. 1988. Public Health in England: The Report of the Committee of Inquiry into the Future Development to the Public Health Function, London, The Stationery Office.

ADAMS, J., VAN DER WAAL, Z., RUSHTON, S. & RANKIN, J. 2018. Associations between introduction and withdrawal of a financial incentive and timing of attendance for antenatal care and incidence of small for gestational age: natural experimental evaluation using interrupted time series methods. BMJ Open, 8, e017697.

AI, NORTON, E. C. & YANG 2011. Extending Regression Discontinuity Models Beyond the Jump Point. HEDG, c/o Department of Economics, University of York.

AKL, E. A., KENNEDY, C., KONDA, K., CACERES, C. F., HORVATH, T., AYALA, G., DOUPE, A., GERBASE, A., WIYSONGE, C. S., SEGURA, E. R., SCHUNEMANN, H. J. & LO, Y. R. 2012. Using GRADE methodology for the development of public health guidelines for the prevention and treatment of HIV and other STIs among men who have sex with men and transgender people. BMC Public Health, 12, 386.

ALAM, A. & BAEZ, J. E. 2011. Does cash for school influence young women's behavior in the longer term? evidence from Pakistan. The World Bank, Policy Research Working Paper Series: 5669.

ALBOUY, V. & LEQUIEN, L. 2009. Does compulsory education lower mortality? Journal of Health Economics, 28, 155-168.

ALCOHOL ADVOCACY COALITION. 2017. Changing Scotland’s relationship with alcohol: Recommendations for further action [Online]. Alcohol Focus Scotland. Available: http://www.alcohol-focus-scotland.org.uk/media/222528/Alcohol-strategy-recommendations-Report.pdf [Accessed June 23 2018].

ALEXANDER, P. E., BERO, L., MONTORI, V. M., BRITO, J. P., STOLTZFUS, R., DJULBEGOVIC, B., NEUMANN, I., RAVE, S. & GUYATT, G. 2014. World Health Organization recommendations are often strong based on low confidence in effect estimates. Journal of Clinical Epidemiology, 67, 629-634.

ALEXANDER, P. E., LI, S.-A., GIONFRIDDO, M. R., STOLTZFUS, R. J., NEUMANN, I., BRITO, J. P., DJULBEGOVIC, B., MONTORI, V. M., SCHÜNEMANN, H. J. &

GUYATT, G. H. 2016. Senior GRADE methodologists encounter challenges as part of WHO guideline development panels: an inductive content analysis. Journal of Clinical Epidemiology, 70, 123-128.

ALI, M. S., GROENWOLD, R. H. H., BELITSER, S. V., PESTMAN, W. R., HOES, A. W., ROES, K. C. B., BOER, A. D. & KLUNGEL, O. H. 2015. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. Journal of Clinical Epidemiology, 68, 122-131.

ALLAM, M. F. 2016. Breast cancer and deodorants/antiperspirants: a systematic review. Central European Journal of Public Health, 24, 245-247.

ALMOND, D., DOYLE, J. J., JR., KOWALSKI, A. E. & WILLIAMS, H. 2010. Estimating marginal returns to medical care: Evidence from at-risk newborns. Quarterly Journal of Economics, 125, 591-634.

ALMOND, D. & DOYLE JR, J. J. 2011. After midnight: A regression discontinuity design in length of postpartum hospital stays. American Economic Journal: Economic Policy, 3, 1-34.

AMETHYST INITIATIVE. Amethyst Initiative Statement [Online]. Available: http://www.theamethystinitiative.org/statement/ [Accessed June 30 2018].

ANDALÕN, M. 2011. Oportunidades to reduce overweight and obesity in Mexico? Health Economics, 20, 1-18.

ANDERSON, M., DOBKIN, C. & GROSS, T. 2012. The effect of health insurance coverage on the use of medical services. American Economic Journal: Economic Policy, 4, 1-27.

ANDERSON, M. L., DOBKIN, C. & GROSS, T. 2014. The effect of health insurance on emergency department visits: Evidence from an age-based eligibility threshold. Review of Economics and Statistics, 96, 189-195.

ANDERSON, P., CHISHOLM, D. & FUHR, D. C. 2009. Effectiveness and cost-effectiveness of policies and programmes to reduce the harm caused by alcohol. The Lancet, 373, 2234-2246.

ANDERSON, P. M., BUTCHER, K. F., CASCIO, E. U. & SCHANZENBACH, D. W. 2011. Is being in school better? The impact of school on children's BMI when starting age is endogenous. Journal of Health Economics, 30, 977-986.

ANGRIST, J. D. & KRUEGER, A. B. 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives, 15, 69-85.

ANGRIST, J. D. & PISCHKE, J.-S. 2009. Mostly harmless econometrics: an empiricist's companion, Princeton and Oxford, Princeton University Press.

ANGRIST, J. D. & PISCHKE, J.-S. 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. The Journal of Economic Perspectives, 24, 3-30.

ANGRIST, J. D. & PISCHKE, J.-S. 2015. Mastering metrics, Princeton and Oxford, Princeton University Press.

ARCAND, J. L. & WOUABE, E. D. 2010. Teacher training and HIV/AIDS prevention in West Africa: Regression discontinuity design evidence from the Cameroon. Health Economics, 19, 36-54.

ARMIJO-OLIVO, S., OSPINA, M., DA COSTA, B. R., EGGER, M., SALTAJI, H., FUENTES, J., HA, C. & CUMMINGS, G. G. 2014. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane Risk of Bias Tool in physical therapy trials. PLOS ONE, 9, e96920.

ARMIJO-OLIVO, S., STILES, C. R., HAGEN, N. A., BIONDO, P. D. & CUMMINGS, G. G. 2012. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. Journal of Evaluation in Clinical Practice, 18, 12-18.

ARMSTRONG, R., WATERS, E., MOORE, L., DOBBINS, M., PETTMAN, T., BURNS, C., SWINBURN, B., ANDERSON, L. & PETTICREW, M. 2014. Understanding evidence: a statewide survey to explore evidence-informed public health decision-making in a local government setting. Implementation Science, 9, 188.

AUSTIN, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46, 399-424.

BABOR THOMAS, F. 2008. Alcohol research and the alcoholic beverage industry: issues, concerns and conflicts of interest. Addiction, 104, 34-47.

BAIOCCHI, M., SMALL, D. S., YANG, L., POLSKY, D. & GROENEVELD, P. W. 2012. Near/far matching: a study design approach to instrumental variables. Health Services and Outcomes Research Methodology, 12, 237-253.

BALLARD, M. & MONTGOMERY, P. 2017. Risk of bias in overviews of reviews: a scoping review of methodological guidance and four-item checklist. Research Synthesis Methods, 8, 92-108.

BANKS, J. & MAZZONNA, F. 2012. The effect of education on old age cognitive abilities: Evidence from a regression discontinuity design. Economic Journal, 122, 418-448.

BARGERLUX, M. J., HEANEY, R. P. & DAVIES, K. M. 1992. Use of the regression discontinuity design in a clinical-trial of calcium efficacy. Journal of Bone and Mineral Research, 7, S189-S189.

BARTHOLOMEW, L. K., PARCEL, G. S. & KOK, G. 1998. Intervention mapping: A process for developing theory- and evidence-based health education programs. Health Education and Behavior, 25, 545.

BAUHOFF, S., HOTCHKISS, D. R. & SMITH, O. 2011. The impact of medical insurance for the poor in Georgia: A regression discontinuity approach. Health Economics, 20, 1362-1378.

BEAGLEHOLE, R. & BONITA, R. 2009. Alcohol: a global health priority. The Lancet, 373, 2173-2174.

BEAL, S. J. & KUPZYK, K. A. 2014. An introduction to propensity scores: what, when, and how. The Journal of Early Adolescence, 34, 66-92.

BECKER, N. V. 2018. The impact of insurance coverage on utilization of prescription contraceptives: evidence from the Affordable Care Act. Journal of Policy Analysis and Management, 37, 571-601.

BEHRMAN, J. A. 2015. The effect of increased primary schooling on adult women's HIV status in Malawi and Uganda: Universal Primary Education as a natural experiment. Social Science and Medicine, 127, 108-115.

BELENKIY, M. 2010. Essays in Applied Microeconomics. Ph.D., University of California, Santa Cruz.

BENACH, J., MALMUSI, D., YASUI, Y. & MARTÍNEZ, J. M. 2013. A new typology of policies to tackle health inequalities and scenarios of impact based on Rose's population approach. Journal of Epidemiology and Community Health, 67, 286-291.

BENCHIMOL, E. I., SMEETH, L., GUTTMANN, A., HARRON, K., MOHER, D., PETERSEN, I., SØRENSEN, H. T., VON ELM, E., LANGAN, S. M. & COMMITTEE, R. W. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Medicine, 12, e1001885.

BENJAMIN, S., MASAI, E., KAMIMURA, N., TAKAHASHI, K., ANDERSON, R. C. & FAISAL, P. A. 2017. Phthalates impact human health: Epidemiological evidences and plausible mechanism of action. Journal of Hazardous Materials, 340, 360-383.

BERNAL, N., CARPIO, M. A. & KLEIN, T. J. 2014. The effects of access to health insurance for informally employed individuals in Peru. Institute for the Study of Labor (IZA).

BEUCHERT, L. V., HUMLUM, M. K. & VEJLIN, R. 2014. The Length of Maternity Leave and Family Health. School of Economics and Management, University of Aarhus, Economics Working Papers.

BEUERMANN, D. W. 2010. The effect of health insurance on health care utilization: evidence from the Medical Expenditure Panel Survey 2000-2005. Journal of CENTRUM Cathedra, 3, 18-31.

BHALOTRA, S., CLOTS-FIGUERAS, I., CASSAN, G. & IYER, L. 2014. Religion, politician identity and development outcomes: Evidence from India. Journal of Economic Behavior and Organization, 104, 4-17.

BHARADWAJ, P., VELLESEN LØKEN, K. & NEILSON, C. 2013. Early life health interventions and academic achievement. American Economic Review, 103, 1862-91.

BOBONIS, G., GONZALEZ-BRENES, M. & CASTRO, R. 2006. Female income, women's status, and spousal violence: effects of the Mexican Oportunidades Program. [Unpublished] 2006. Presented at the Population Association of America, 2006 Annual Meeting, Los Angeles, California, March 30 - April 1, 2006.

BOES, S. & STILLMAN, S. 2013. Does changing the legal drinking age influence youth behaviour? : Institute for the Study of Labor (IZA).

BOHEIM, R. & LEONI, T. 2014. Firms' sickness costs and workers' sickness absences. National Bureau of Economic Research, Inc, NBER Working Papers: 20305.

BONELL, C. P., HARGREAVES, J., COUSENS, S., ROSS, D., HAYES, R., PETTICREW, M. & KIRKWOOD, B. R. 2011. Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. Journal of Epidemiology and Community Health, 65, 582-587.

BONIFACE, S., SCANNELL, J. W. & MARLOW, S. 2017. Evidence for the effectiveness of minimum pricing of alcohol: a systematic review and assessment using the Bradford Hill criteria for causality. BMJ Open, 7, e013497.

BOR, J. 2013. Essays in the Economics of HIV/AIDS in Rural South Africa. PhD, Harvard.

BOR, J., MOSCOE, E. & BARNIGHAUSEN, T. 2015. Three approaches to causal inference in regression discontinuity designs. Epidemiology, 26, E28-E30.

BOR, J., MOSCOE, E., MUTEVEDZI, P., NEWELL, M. L. & BARNIGHAUSEN, T. 2014. Regression discontinuity designs in epidemiology: causal inference without randomized trials. Epidemiology, 25, 729-37.

BORGERSON, K. 2009. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. Perspectives in Biology and Medicine, 52, 218-33.

BREAST CANCER UK. 2017. Endocrine disrupting chemicals [Online]. Available: https://www.breastcanceruk.org.uk/science-and-research/background-briefings/endocrine-disrupting-chemicals/ [Accessed September 13 2018].

BREEZE, P., WOMACK, R., PRYCE, R., BRENNAN, A. & GOYDER, E. 2018. The impact of a local sugar sweetened beverage health promotion and price increase on sales in public leisure centre facilities. PLoS ONE, 13, e0194637.

BRODY, J. G., MOYSICH, K. B., HUMBLET, O., ATTFIELD, K. R., BEEHLER, G. P. & RUDEL, R. A. 2007. Environmental pollutants and breast cancer: epidemiologic studies. Cancer, 109, 2667-711.

BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. & STÜRMER, T. 2006. Variable selection for propensity score models. American Journal of Epidemiology, 163, 1149-1156.

BROWN, C. A. & LILFORD, R. J. 2006. The stepped wedge trial design: a systematic review. BMC Medical Research Methodology, 6, 54-54.

BROWNSON, R. C., BAKER, E. A., LEET, T. L., GILLESPIE, K. N. & TRUE, W. R. 2011. Evidence-based public health, Oxford, Oxford University Press.

BUCHMUELLER, T. C., GRAZIER, K., HIRTH, R. A. & OKEKE, E. N. 2013. The price sensitivity of Medicare beneficiaries: a regression discontinuity approach. Health Economics, 22, 35-51.

BUNCE, M. M. 2008. Pay-for-performance's impact on overall quality of care for acute myocardial infarction patients. 68, ProQuest Information & Learning.

BURGER, N. E., KAFFINE, D. T. & YU, B. 2014. Did California's hand-held cell phone ban reduce accidents? Transportation Research Part A: Policy and Practice, 66, 162-172.

BURNS, M. E., DAGUE, L., DELEIRE, T., DORSCH, M., FRIEDSAM, D., LEININGER, L. J., PALMUCCI, G., SCHMELZER, J. & VOSKUIL, K. 2014. The effects of expanding public insurance to rural low-income childless adults. Health Services Research, 49, 2173-2187.

BURTON, R., HENN, C., LAVOIE, D., O'CONNOR, R., PERKINS, C., SWEENEY, K., GREAVES, F., FERGUSON, B., BEYNON, C., BELLONI, A., MUSTO, V., MARSDEN, J.

& SHERON, N. 2017. A rapid evidence review of the effectiveness and cost-effectiveness of alcohol control policies: an English perspective. The Lancet, 389, 1558-1580.

CALLAGHAN, R. C., GATLEY, J. M., SANCHES, M. & ASBRIDGE, M. 2014b. Impacts of the minimum legal drinking age on motor vehicle collisions in Québec, 2000-2012. American Journal of Preventive Medicine, 47, 788-795.

CALLAGHAN, R. C., GATLEY, J. M., SANCHES, M., ASBRIDGE, M. & STOCKWELL, T. 2016b. Impacts of drinking-age legislation on alcohol-impaired driving crimes among young people in Canada, 2009-13. Addiction, 111, 994-1003.

CALLAGHAN, R. C., GATLEY, J. M., SANCHES, M. & BENNY, C. 2016a. Do drinking-age laws have an impact on crime? Evidence from Canada, 2009-2013. Drug and alcohol dependence, 167, 67-74.

CALLAGHAN, R. C., GATLEY, J. M., SANCHES, M., BENNY, C. & ASBRIDGE, M. 2016c. Release from drinking-age restrictions is associated with increases in alcohol-related motor vehicle collisions among young drivers in Canada. Preventive Medicine, 91, 356-363.

CALLAGHAN, R. C., SANCHES, M. & GATLEY, J. M. 2013a. Impacts of the minimum legal drinking age legislation on in-patient morbidity in Canada, 1997-2007: A regression-discontinuity approach. Addiction, 108, 1590-1600.

CALLAGHAN, R. C., SANCHES, M., GATLEY, J. M. & CUNNINGHAM, J. K. 2013b. Effects of the minimum legal drinking age on alcohol-related health service use in hospital settings in Ontario: A regression-discontinuity approach. American Journal of Public Health, 103, 2284-2291.

CALLAGHAN, R. C., SANCHES, M., GATLEY, J. M. & STOCKWELL, T. 2014a. Impacts of drinking-age laws on mortality in Canada, 1980-2009. Drug and Alcohol Dependence, 138, 137-145.

CAMACHO, A. & CONOVER, E. 2013. Effects of subsidized health insurance on newborn health in a developing country. Economic development and cultural change, 61, 633-658.

CAMPBELL, D. T. & ROSS, H. L. 1968. The Connecticut crackdown on speeding: time-series data in quasi-experimental analysis. Law & Society Review, 3, 33-54.

CAMPOSTRINI, S., HOLTZMAN, D., MCQUEEN, D. V. & BOARETTO, E. 2006. Evaluating the effectiveness of health promotion policy: changes in the law on drinking and driving in California. Health Promotion International, 21, 130-135.

CANCER RESEARCH UK. 2016. Hormones in our environment [Online]. Available: https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/hormones-and-cancer/hormones-in-our-environment [Accessed September 8 2018].

CAPACCI, S., MAZZOCCHI, M. & SHANKAR, B. 2012. Evaluation with inadequate data: the impact of the French vending machine ban. Agricultural and Applied Economics Association.

CARD, D., DOBKIN, C. & MAESTAS, N. 2008. The impact of nearly universal insurance coverage on health care utilization: evidence from Medicare. The American Economic Review, 98, 2242-2258.

CARD, D., DOBKIN, C. & MAESTAS, N. 2009. Does Medicare save lives? Quarterly Journal of Economics, 124, 597-636.

CARD, D. & SHORE-SHEPPARD, L. D. 2004. Using discontinuous eligibility rules to identify the effects of the federal Medicaid expansions on low-income children. Review of Economics and Statistics, 86, 752-766.

CARDELLA, E. & DEPEW, B. 2014. The effect of health insurance coverage on the reported health of young adults. Economics Letters, 124, 406-410.

CARNEIRO, P., GALASSO, E. & GINJA, R. 2014. Tackling Social Exclusion: Evidence from Chile. C.E.P.R. Discussion Papers, CEPR Discussion Papers: 9950.

CARNEIRO, P. & GINJA, R. 2014. Long-term impacts of compensatory preschool on health and behavior: evidence from Head Start. American Economic Journal-Economic Policy, 6, 135-173.

CARPENTER, C. & DOBKIN, C. 2009. The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age. American Economic Journal: Applied Economics, 1, 164-182.

CARPENTER, C. & DOBKIN, C. 2011. The minimum legal drinking age and public health. Journal of Economic Perspectives, 25, 133-56.

CARPENTER, C. & DOBKIN, C. 2015. The minimum legal drinking age and morbidity in the US [Online]. School of Economic Sciences Faculty and Graduate Student Seminar Series. Washington State University. Available: http://ses.wsu.edu/seminars/ [Accessed 1 December 2017].

CARPENTER, C., DOBKIN, C. & WARMAN, C. 2014. The mechanisms of alcohol control. IZA Discussion Paper Series No. 8720 [Online]. Available: http://ftp.iza.org/dp8720.pdf [Accessed 1 December 2017].

CARRANZA BARONA, C. & MÉNDEZ SAYAGO, J. A. 2015. ¿Mejora el bono de desarrollo humano la lactancia materna exclusiva en Ecuador? (Spanish). Does human

development bonus improve exclusive breastfeeding in Ecuador? (English), 23, 63-81.

CASSWELL, S. & THAMARANGSI, T. 2009. Reducing harm from alcohol: call to action. The Lancet, 373, 2247-2257.

CATS CONSORTIUM 2010. Implementation of CBT for youth affected by the World Trade Center disaster: matching need to treatment intensity and reducing trauma symptoms. Journal of Traumatic Stress, 23, 699-707.

CATTANEO, M. D., KEELE, L., TITIUNIK, R. & VAZQUEZ-BARE, G. 2016. Interpreting regression discontinuity designs with multiple cutoffs. The Journal of Politics, 78, 1229-1248.

CENTERS FOR DISEASE CONTROL AND PREVENTION. 2017. Teen drivers: get the facts [Online]. [Accessed June 30 2018].

CERDA, M., MORENOFF, J. D., HANSEN, B. B., TESSARI HICKS, K. J., DUQUE, L. F., RESTREPO, A. & DIEZ-ROUX, A. V. 2012. Reducing violence by transforming neighborhoods: a natural experiment in Medellin, Colombia. American Journal of Epidemiology, 175, 1045-53.

CHAY, K. Y. & GREENSTONE, M. 2003. Air Quality, Infant Mortality, and the Clean Air Act of 1970. NBER Working Paper No. 10053 [Online]. Available: http://www.nber.org/papers/w10053 [Accessed 24 August 2015].

CHAY, K. Y., KIM, D. & SWAMINATHAN, S. 2010. Medicare, Hospital Utilization and Mortality: Evidence from the Program's Origins [Online]. Available: https://www.chicagofed.org/~/media/others/events/2010/health-care-conference/paper-chay-pdf.pdf [Accessed 2015 December 29.

CHEN, Y., EBENSTEIN, A., GREENSTONE, M. & LI, H. 2013. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. Proceedings of the National Academy of Sciences, 110, 12936-12941.

CHENG, S.-H., LEE, T.-T. & CHEN, C.-C. 2012. A longitudinal examination of a pay-for-performance program for diabetes care: evidence from a natural experiment. Medical Care, 50, 109-116.

CHU, R., WALTER, S. D., GUYATT, G., DEVEREAUX, P. J., WALSH, M., THORLUND, K. & THABANE, L. 2012. Assessment and implication of prognostic imbalance in randomized controlled trials with a binary outcome – a simulation study. PLoS ONE, 7, e36677.

CLARK, D. & ROYER, H. 2013. The effect of education on adult mortality and health: Evidence from Britain. American Economic Review, 103, 2087-2120.

CLARK, M. H. & SHADISH, W. R. 2008. Can nonrandomized experiments yield accurate answers?: a randomized experiment comparing random and nonradom assignments. Journal of the American Statistical Association, 103, 1334-1356.

COBURN, J. L. 2009. The Effect of Tennessee's Prekindergarten Programs on Young Children's School Readiness Skills: A Regression Discontinuity Design. Ph.D., Tennessee Technological University.

COGNEAU, D., MESPLE-SOMPS, S. & SPIELVOGEL, G. 2013. Development at the border: policies and national integration in Cote d'Ivoire and its neighbors. The World Bank, Policy Research Working Paper Series: 6626.

CONLEY, D. & HEERWIG, J. 2012. The long-term effects of military conscription on mortality: estimates from the Vietnam-era draft lottery. Demography, 49, 841-855.

CONOVER, E. & SCRIMGEOUR, D. 2013. Health consequences of easier access to alcohol: New Zealand evidence. Journal of Health Economics, 32, 570-585.

CONSORTIUM, C. 2010. Implementation of CBT for youth affected by the World Trade Center disaster: matching need to treatment intensity and reducing trauma symptoms. Journal of Traumatic Stress, 23, 699-707.

COOK, T. D. 2008. "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics, 142, 636-654.

COOK, T. D. 2018. Twenty-six assumptions that have to be met if single random assignment experiments are to warrant “gold standard” status: A commentary on Deaton and Cartwright. Social Science & Medicine, 210, 37-40.

COOK, T. D. & WONG, V. C. 2008. Empirical tests of the validity of the regression discontinuity design. Annales d'économie et de statistique, 91/92, 127-150.

COUDIN, E., PLA, A. & SAMSON, A.-L. 2014. GPs' response to price regulation: evidence from a nationwide French reform. Centre de Recherche en Economie et Statistique, Working Papers: 2014-14.

COUSENS, S., HARGREAVES, J., BONELL, C., ARMSTRONG, B., THOMAS, J., KIRKWOOD, B. R. & HAYES, R. 2011. Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. Journal of Epidemiology and Community Health, 65, 576-581.

CRAIG, P., COOPER, C., GUNNELL, D., HAW, S., LAWSON, K., MACINTYRE, S., OGILVIE, D., PETTICREW, M., REEVES, B., SUTTON, M. & THOMPSON, S. 2011. Using natural experiments to evaluate population health interventions: guidance for

producers and users of evidence [Online]. Available: http://www.mrc.ac.uk/documents/pdf/natural-experiments-guidance/ [Accessed 5 December 2014].

CRAIG, P., COOPER, C., GUNNELL, D., HAW, S., LAWSON, K., MACINTYRE, S., OGILVIE, D., PETTICREW, M., REEVES, B., SUTTON, M. & THOMPSON, S. 2012. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. Journal of Epidemiology and Community Health, 66, 1182-1186.

CRAIG, P., GIBSON, M., CAMPBELL, M., POPHAM, F. & KATIKIREDDI, S. V. 2018. Making the most of natural experiments: What can studies of the withdrawal of public health interventions offer? Preventive Medicine, 108, 17-22.

CRAIG, P., KATIKIREDDI, S. V., LEYLAND, A. & POPHAM, F. 2017. Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. Annual Reviews in Public Health, 38, 39-56.

CROST, B., FELTER, J. & JOHNSTON, P. 2014. Aid under fire: Development projects and civil conflict. American Economic Review, 104, 1833-1856.

CROST, B. & GUERRERO, S. 2012. The effect of alcohol availability on marijuana use: Evidence from the minimum legal drinking age. Journal of Health Economics, 31, 112-121.

CROST, B. & REES, D. I. 2013. The minimum legal drinking age and marijuana use: New estimates from the NLSY97. Journal of Health Economics, 32, 474-476.

CULLEN, K. W., KOEHLY, L. M., ANDERSON, C., BARANOWSKI, T., PROKHOROV, A., BASEN-ENGQUIST, K., WETTER, D. & HERGENROEDER, A. 1999. Gender differences in chronic disease risk behaviors through the transition out of high school. American Journal of Preventive Medicine, 17, 1-7.

CULLEN, K. W., KOEHLY, L. M., ANDERSON, C., BARANOWSKI, T., PROKHOROV, A., BASEN-ENGQUIST, K., WETTER, D. & HERGENROEDER, A. 1999. Gender differences in chronic disease risk behaviors through the transition out of high school. American Journal of Preventive Medicine, 17, 1-7.

CUTTER, W. B. & NEIDELL, M. 2009. Voluntary information programs and environmental regulation: Evidence from 'Spare the Air'. Journal of Environmental Economics and Management, 58, 253-265.

DAGUE, L. 2014. The effect of Medicaid premiums on enrollment: A regression discontinuity approach. Journal of Health Economics, 37, 1-12.

DANIELS, V., SOMERS, M., ORFORD, J. & KIRBY, B. 1992. How can risk drinking amongst medical patients be modified? The effects of computer screening and advice and a self-help manual. Behavioural Psychotherapy, 20, 47-60.

D'ASCENZO, F., CAVALLERO, E., BIONDI-ZOCCAI, G., MORETTI, C., OMEDE, P., BOLLATI, M., CASTAGNO, D., MODENA, M. G., GAITA, F. & SHEIBAN, I. 2012. Use and misuse of multivariable approaches in interventional cardiology studies on drug-eluting stents: a systematic review. Journal of Interventional Cardiology, 25, 611-21.

DAVEY SMITH, G., EBRAHIM, S. & FRANKEL, S. 2001. How policy informs the evidence. BMJ, 322, 184.

DAVIES, N. M., SMITH, G. D., WINDMEIJER, F. & MARTIN, R. M. 2013. Issues in the reporting and conduct of instrumental variable studies: a systematic review. Epidemiology, 24, 363-369.

DAYER, M. J., JONES, S., PRENDERGAST, B., BADDOUR, L. M., LOCKHART, P. B. & THORNHILL, M. H. 2015. Incidence of infective endocarditis in England, 2000-13: a secular trend, interrupted time-series analysis. Lancet, 385, 1219-28.

DAYSAL, N. M., TRANDAFIR, M. & VAN EWIJK, R. 2013. Returns to childbirth technologies: evidence from preterm births. Institute for the Study of Labor (IZA).

DE BRAUW, A. 2012. Regression discontinuity impacts with an implicit index: evaluating El Salvador's Comunidades Solidarias Rurales Transfer Programme. International Policy Centre for Inclusive Growth.

DE BRAUW, A. & PETERMAN, A. 2011. Can conditional cash transfers improve maternal health and birth outcomes?: Evidence from El Salvador's Comunidades Solidarias Rurales. IFPRI Discussion Paper No. 1080. International Food Policy Research Institute (IFPRI).

DE LA MATA, D. 2012. The effect of Medicaid eligibility on coverage, utilization, and children's health. Health Economics, 21, 1061-79.

DE PAOLA, M., SCOPPA, V. & FALCONE, M. 2013. The deterrent effects of the penalty points system for driving offences: A regression discontinuity approach. Empirical Economics, 45, 965-985.

DEATON, A. & CARTWRIGHT, N. 2018. Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2-21.

DECKER, S. L. 2005. Medicare and the health of women with breast cancer. Journal of Human Resources, 40, 948-968.

DEEKS, J., DINNES, J., D'AMICO, R., SOWDEN, A. J., SAKAROVITCH, C., SONG, F., PETTICREW, M. & ALTMAN, D. G. 2003. Evaluating non-randomised intervention studies. Health Technology Assessment, 7, 1-192.

DEJONG, W. & BLANCHETTE, J. 2014. Case closed: research evidence on the positive public health impact of the age 21 minimum legal drinking age in the United States. Journal of Studies on Alcohol and Drugs, 75 Suppl 17, 108-15.

DEJONG, W. & BLANCHETTE, J. 2014. When enough is enough: the public health argument for the age 21 minimum legal drinking age. Journal of Studies on Alcohol and Drugs, 75, 1050-1052.

DEL BONO, E., FRANCESCONI, M. & BEST, N. G. 2011. Health information and health outcomes: an application of the regression discontinuity design to the 1995 UK contraceptive pill scare case. Institute for Social and Economic Research.

DELL, M. 2010. The persistent effects of Peru's mining mita. Econometrica, 78, 1863-1903.

DEVITT, T. S. 2006. Policy change regarding substance abuse in integrated dual disorders residential treatment. Ph.D., The Adler School of Professional Psychology (Chicago).

DEZA, M. 2013. Essays on Drug Use and Crime. PhD., University of California, Berkeley.

DEZA, M. 2015. The effects of alcohol on the consumption of hard drugs: regression discontinuity evidence from the national longitudinal study of youth, 1997. Health Economics, 24, 419-38.

DICKERT-CONLIN, S. & ELDER, T. 2010. Suburban legend: School cutoff dates and the timing of births. Economics of Education Review, 29, 826-841.

DILL, J., MCNEIL, N., BROACH, J. & MA, L. 2014. Bicycle boulevards and changes in physical activity and active transportation: Findings from a natural experiment. Preventive Medicine, 69 Suppl 1, S74-8.

DISMEVAL CONSORTIUM 2012. DISMEVAL : developing and validating disease management evaluation methods for European healthcare systems RAND Corporation Technical Reports. Santa Monica, CA: RAND Corporation.

DOMBROWSKI, S. U., SNIEHOTTA, F. F., AVENELL, A., JOHNSTON, M., MACLENNAN, G. & ARAÚJO-SOARES, V. 2012. Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. Health Psychology Review, 6, 7-32.

DUGAN, J., VIRANI, S. S. & HO, V. 2012. Medicare eligibility and physician utilization among adults with coronary heart disease and stroke. Medical Care, 50, 547-553.

DUNNING, T. 2012. Natural experiments in the social sciences: a design-based approach, Cambridge, Cambridge University Press.

EFFECTIVE PUBLIC HEALTH PRACTICE PROJECT. 2010. Quality Assessment Tool for Quantitative Studies [Online]. Available: http://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf [Accessed 29 June 2015].

EGGER, M., EGGER, M., SMITH, G. D. & ALTMAN, D. G. 2001. Systematic reviews in health care: meta-analysis in context, London, BMJ Books.

EIBICH, P. 2014. Understanding the effect of retirement on health using regression discontinuity design. SOEP Papers on Multidisciplinary Panel Data Research. Berlin: German Institute for Economic Research (DIW Berlin).

ELDER, T. E. 2010. The importance of relative standards in ADHD diagnoses: evidence based on exact birth dates. Journal of Health Economics, 29, 641-656.

ELLEN, M. E., LÉON, G., BOUCHARD, G., OUIMET, M., GRIMSHAW, J. M. & LAVIS, J. N. 2014. Barriers, facilitators and views about next steps to implementing supports for evidence-informed decision-making in health systems: a qualitative study. Implementation Science, 9, 179.

ENGMANN, N. J., GOLMAKANI, M. K., MIGLIORETTI, D. L., SPRAGUE, B. L., KERLIKOWSKE, K. & FOR THE BREAST CANCER SURVEILLANCE, C. 2017. Population-attributable risk proportion of clinical risk factors for breast cancer. JAMA Oncology, 3, 1228-1236.

ERTAN YORUK, C. 2012. Three Essays on the Impacts of Public Policies on Labor Market and Health Outcomes. Ph.D., Northeastern University.

ERTAN YORUK, C. & YORUK, B. K. 2015. Alcohol consumption and risky sexual behavior among young adults: evidence from minimum legal drinking age laws. Journal of Population Economics, 28, 133-157.

ERTAN YÖRÜK, C. & YÖRÜK, B. K. 2012. The impact of drinking on psychological well-being: Evidence from minimum drinking age laws in the United States. Social Science and Medicine, 75, 1844-1854.

ESPELT, A., VILLALBI, J. R., BOSQUE-PROUS, M., PARES-BADELL, O., MARI-DELL'OLMO, M. & BRUGAL, M. T. 2017. The impact of harm reduction programs and police interventions on the number of syringes collected from public spaces. A time

series analysis in Barcelona, 2004-2014. International Journal of Drug Policy, 50, 11-18.

ESPINOSA, S. 2014. Intended and unintended incentives in social protection programmes: evidence from Colombia and Mexico. Ph.D., University College London.

EVANS, M. E., BANKS, S. M., HUZ, S. & MCNULTY, T. L. 1994. Initial hospitalization and community tenure outcomes of intensive case management for children and youth with serious emotional disturbance. Journal of Child & Family Studies, 3, 225-234.

EVANS, W. N., MORRILL, M. S. & PARENTE, S. T. 2010. Measuring inappropriate medical diagnosis and treatment in survey data: The case of ADHD among school-age children. Journal of Health Economics, 29, 657-673.

FÉ, E. & HOLLINGSWORTH, B. 2012. Estimating the effect of retirement on mental health via panel discontinuity designs. University Library of Munich, Germany.

FILMER, D. & SCHADY, N. 2014. The medium-term effects of scholarships in a low-income country. Journal of Human Resources, 49, 663-694.

FISCHER, A. J., THRELFALL, A., MEAH, S., COOKSON, R., RUTTER, H. & KELLY, M. P. 2013. The appraisal of public health interventions: an overview. Journal of Public Health (Oxford, England), 35, 488-494.

FISHER, R. A. 1935. The design of experiments, 8th ed. (1966; reprinted 1971), New York, Hafner.

FLAM-ZALCMAN, R., MANN, R. E., STODUTO, G., NOCHAJSKI, T. H., RUSH, B. R., KOSKI-JÄNNES, A., WICKENS, C. M., THOMAS, R. K. & REHM, J. 2013. Evidence from regression-discontinuity analyses for beneficial effects of a criterion-based increase in alcohol treatment. International Journal of Methods in Psychiatric Research, 22, 59-70.

FLETCHER, J. M. 2014. Enhancing the gene-environment interaction framework through a quasi-experimental research design: Evidence from differential responses to September 11. Biodemography and Social Biology, 60, 1-20.

FORESIGHT 2007. Tackling obesities: future choices. 2nd ed. London: Government Office for Science.

FOXCROFT, D. R. & TSERTSVADZE, A. 2011. Universal multi-component prevention programs for alcohol misuse in young people. Cochrane Database of Systematic Reviews.

FOXCROFT, D. R. & TSERTSVADZE, A. 2011. Universal school-based prevention programs for alcohol misuse in young people. Cochrane Database of Systematic Reviews.

FREEDMAN, D. A. 2010. Statistical models and causal inference : a dialogue with the social sciences, Cambridge, Cambridge University Press.

FREEDMAN, D. A. & BERK, R. A. 2008. Weighting regressions by propensity scores. Evaluation Review, 32, 392-409.

FU, Z., ZHAO, F., CHEN, K., XU, J., LI, P., XIA, D. & WU, Y. 2017. Association between urinary phthalate metabolites and risk of breast cancer and uterine leiomyoma. Reproductive Toxicology, 74, 134-142.

GARABEDIAN, L. F., CHU, P., TOH, S., ZASLAVSKY, A. M. & SOUMERAI, S. B. 2014. Potential bias of instrumental variable analyses for observational comparative effectiveness research. Annals of Internal Medicine, 161, 131-8.

GARCIA-GOMEZ, P. & GIELEN, A. C. 2014. Health Effects of Containing Moral Hazard: Evidence from Disability Insurance Reform. Tinbergen Institute, Tinbergen Institute Discussion Papers: 14-102/V.

GARROUSTE, C., LE, J. & MAURIN, E. 2011. The choice of detecting Down syndrome: Does money matter? Health Economics, 20, 1073-1089.

GAYAT, E., PIRRACCHIO, R., RESCHE-RIGON, M., MEBAZAA, A., MARY, J. Y. & PORCHER, R. 2010. Propensity scores in intensive care and anaesthesiology literature: a systematic review. Intensive Care Medicine, 36, 1993-2003.

GELMAN, A. 2018. Benefits and limitations of randomized controlled trials: A commentary on Deaton and Cartwright. Social Science & Medicine, 210, 48-49.

GERA, R., MOKBEL, R., IGOR, I. & MOKBEL, K. 2018. Does the use of hair dyes increase the risk of developing breast cancer? a meta-analysis and review of the literature. Anticancer Research, 38, 707-716.

GERARD, J. 2007. Should we raise the age of legal drinking? Public Policy Research, 14, 31-35.

GIOVANIS, E. 2015. The effect of smog-ozone warnings and a vanpool program on traffic volume in York County of South Carolina. Environment and Planning B-Planning & Design, 42, 195-220.

GLANCE, L. G., OSLER, T. M., MUKAMEL, D. B., MEREDITH, J. W. & DICK, A. W. 2014. Effectiveness of nonpublic report cards for reducing trauma mortality. JAMA Surgery, 149, 137-143.

GONZÁLEZ, L. 2013. The effect of a universal child benefit on conceptions, abortions, and early maternal labor supply. American Economic Journal: Economic Policy, 5, 160-188.

GORDON, D. & MILLER, D. L. 2012. The South African pension program and the health of the elderly and their families: regression discontinuity evidence from October Household Surveys [Online]. Available: http://faculty.econ.ucdavis.edu/faculty/dlmiller/research/papers/Gordon_Miller_2012-06-08.pdf.

GORE, A. C., CHAPPELL, V. A., FENTON, S. E., FLAWS, J. A., NADAL, A., PRINS, G. S., TOPPARI, J. & ZOELLER, R. T. 2015. EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. Endocrine Reviews, 36, 1-150.

GORMLEY, W. T., JR., GAYER, T. & PHILLIPS, D. 2005. The effects of universal Pre-K on cognitive development. Developmental Psychology, 41, 872-884.

GOUGH, D., THOMAS, J. & OLIVER, S. 2012. Clarifying differences between review designs and methods. Systematic reviews, 1, 28.

GOVERNMENT OF CANADA. 2008. Bisphenol A fact sheet [Online]. Available: https://www.canada.ca/en/health-canada/services/chemical-substances/fact-sheets/chemicals-glance/bisphenol-a.html [Accessed September 13 2018].

GRAY, J. M., RASANAYAGAM, S., ENGEL, C. & RIZZO, J. 2017. State of the evidence 2017: an update on the connection between breast cancer and the environment. Environmental Health: A Global Access Science Source, 16, 94.

GREENHALGH, T., HOWICK, J. & MASKREY, N. 2014. Evidence based medicine: a movement in crisis? BMJ, 348, g3725.

GREENLAND, S. 2000. An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology, 29, 722-729.

GREENWOOD, E. 2012. Essays in Social Economics. Ph.D., Harvard University.

GUERTZGEN, N. & HANK, K. 2014. Maternity leave and mothers' long-term sickness absence: Evidence from Germany. ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research.

GUO, S. & FRASER, M. W. 2010. Propensity score analysis: statistical methods and applications, London, Sage.

GUPTA, N. & NIELSSON, J. 2017. Short- and long-term effects of adolescent alcohol access: evidence from Denmark. Economics Working Papers. Aarhus University.

GUTHMULLER, S. & WITTWER, J. 2012. L'effet de la Couverture maladie universelle complémentaire (CMU-C) sur le nombre de visites chez le médecin: une analyse par régression sur discontinuités. Economie publique, 1-2, 71-94.

HANANDITA, W. & TAMPUBOLON, G. 2014. Does poverty reduce mental health? An instrumental variable analysis. Social Science & Medicine, 113, 59-67.

HANBURY, A., FARLEY, K., THOMPSON, C., WILSON, P. M., CHAMBERS, D. & HOLMES, H. 2013. Immediate versus sustained effects: interrupted time series analysis of a tailored intervention. Implementation Science, 8, 130-130.

HANLON, P., CARLISLE, S., HANNAH, M. & LYON, A. 2012. The future public health, Maidenhead, Open University Press.

HANSEN, B. 2014. Punishment and deterrence: evidence from drunk driving. National Bureau of Economic Research, Inc, NBER Working Papers: 20243.

HANSEN, B. 2015. Punishment and deterrence: Evidence from drunk driving. American Economic Review, 105, 1581-1617.

HARDEFELDT, P. J., EDIRIMANNE, S. & ESLICK, G. D. 2013. Deodorant use and breast cancer risk. Epidemiology, 24, 172.

HARDER, T., ABU SIN, M., BOSCH-CAPBLANCH, X., BRUNO, C., DE CARVALHO GOMES, H., DUCLOS, P., ECKMANNS, T., ELDER, R., ELLIS, S., FORLAND, F., GARNER, P., JAMES, R., JANSEN, A., KRAUSE, G., LÉVY-BRUHL, D., MORGAN, A., MEERPOHL, J. J., NORRIS, S., REHFUESS, E., SÁNCHEZ-VIVAR, A., SCHÜNEMANN, H., TAKLA, A., WICHMANN, O., ZINGG, W. & ZUIDERENT-JERAK, T. 2015. Towards a framework for evaluating and grading evidence in public health. Health Policy, 119, 732-736.

HARTLING, L., HAMM, M. P., MILNE, A., VANDERMEER, B., SANTAGUIDA, P. L., ANSARI, M., TSERTSVADZE, A., HEMPEL, S., SHEKELLE, P. & DRYDEN, D. M. 2013. Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. Journal of Clinical Epidemiology, 66, 973-981.

HARTLING, L., MILNE, A., HAMM, M. P., VANDERMEER, B., ANSARI, M., TSERTSVADZE, A. & DRYDEN, D. M. 2013. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. Journal of Clinical Epidemiology, 66, 982-993.

HAVASSAY, B. E. & WASSERMAN, D. A. 1991. Prevalence of comorbidity among cocaine users in treatment. Problems of Drug Dependence 1991: Proceeding of the Annual Scientific Meeting (53rd), the Committee on Problems of Drug

Dependence, Inc. Richmond, Virginia: National Inst. on Drug Abuse, Rockville, MD.

HAWTON, K., BERGEN, H., SIMKIN, S., BROCK, A., GRIFFITHS, C., ROMERI, E., SMITH, K. L., KAPUR, N. & GUNNELL, D. 2009. Effect of withdrawal of co-proxamol on prescribing and deaths from drug poisoning in England and Wales: time series analysis. BMJ: British Medical Journal, 339, 435-438.

HAWTON, K., BERGEN, H., SIMKIN, S., DODD, S., POCOCK, P., BERNAL, W., GUNNELL, D. & KAPUR, N. 2013. Long term effect of reduced pack sizes of paracetamol on poisoning deaths and liver transplant activity in England and Wales: interrupted time series analyses. BMJ (Clinical research ed.), 346, f403.

HAWTON, K., BERGEN, H., SIMKIN, S., WELLS, C., KAPUR, N. & GUNNELL, D. 2012. Six-year follow-up of impact of co-proxamol withdrawal in England and Wales on prescribing and deaths: time-series study. PLoS Medicine, 9, e1001213.

HAYNES, L., SERVICE, O., GOLDACRE, B. & TORGERSON, D. J. 2012. Test, learn, adapt: developing public policy with randomised controlled trials [Online]. Cabinet Office Behavioural Insights Team. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf [Accessed 24 January 2015].

HE, F. J., BRINSDEN, H. C. & MACGREGOR, G. A. 2013. UK population salt reduction: An experiment in public health. Lancet, 382, 43-43.

HEALEY, C., RAHMAN, A., FAIZAL, M. & KINDERMAN, P. 2014. Underage drinking in the UK: changing trends, impact and interventions. A rapid evidence synthesis. International Journal on Drug Policy, 25, 124-32.

HENSCHEL, S., ATKINSON, R., ZEKA, A., LE TERTRE, A., ANALITIS, A., KATSOUYANNI, K., CHANEL, O., PASCAL, M., FORSBERG, B., MEDINA, S. & GOODMAN, P. G. 2012. Air pollution interventions and their impact on public health. International Journal of Public Health, 57, 757-768.

HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S., WERLER, M. M. & MITCHELL, A. A. 2002. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology, 155, 176-184.

HERNÁN, M. A. & ROBINS, J. M. 2006. Instruments for causal inference: an epidemiologist's dream? Epidemiology, 17, 360-372.

HIGGINS, J. & GREEN, S. 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [Online]. Available: http://handbook-5-1.cochrane.org/ [Accessed May 20 2018].

HIGGINS, J. P. T., ALTMAN, D. G., GØTZSCHE, P. C., JÜNI, P., MOHER, D., OXMAN, A. D., SAVOVIĆ, J., SCHULZ, K. F., WEEKS, L. & STERNE, J. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ, 343, 889-893.

HØGLEND, P. 1996. Long-term effects of transference interpretations: Comparing results from a quasi-experimental and a naturalistic long-term follow-up study of brief dynamic psychotherapy. Acta Psychiatrica Scandinavica, 93, 205-211.

HØGLEND, P., HEYERDAHL, O., AMLO, S., ENGELSTAD, V., FOSSUM, A., SØRBYE, O. & SØRLIE, T. 1993. Interpretations of the patient-therapist relationship in brief dynamic psychotherapy: effects on long-term mode-specific changes. Journal of Psychotherapy Practice and Research, 2, 296-306.

HOME OFFICE. 2013. Next steps following the consultation on delivering the Government's alcohol strategy [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/223773/Alcohol_consultation_response_report_v3.pdf.

HOU, X. & CHAO, S. 2008. An evaluation of the initial impact of the medical assistance program for the poor in Georgia. The World Bank, Policy Research Working Paper Series: 4588.

HU, T., DECKER, S. L. & CHOU, S.-Y. 2014. The Impact of Health Insurance Expansion on Physician Treatment Choice: Medicare Part D and Physician Prescribing. National Bureau of Economic Research, Inc.

HUANG, W. & ZHOU, Y. 2013. Effects of education on cognition at older ages: evidence from China's Great Famine. Social Science and Medicine, 98, 54-62.

HUCKLE, T. & PARKER, K. 2014. Long-term impact on alcohol-involved crashes of lowering the minimum purchase age in New Zealand. American Journal of Public Health, 104, 1087-91.

HULLEGIE, P. G. J. & KLEIN, T. J. 2010. The effect of private health insurance on medical care utilization and self-assessed health in Germany. Health Economics, 19, 1048-1062.

HULTCRANTZ, M., RIND, D., AKL, E. A., TREWEEK, S., MUSTAFA, R. A., IORIO, A., ALPER, B. S., MEERPOHL, J. J., MURAD, M. H., ANSARI, M. T., KATIKIREDDI, S. V., ÖSTLUND, P., TRANÆUS, S., CHRISTENSEN, R., GARTLEHNER, G., BROZEK, J., IZCOVICH, A., SCHÜNEMANN, H. & GUYATT, G. 2017. The GRADE Working Group clarifies the construct of certainty of evidence. Journal of Clinical Epidemiology, 87, 4-13.

IJAZ, S., VERBEEK, J. H., MISCHKE, C. & RUOTSALAINEN, J. 2014. Inclusion of nonrandomized studies in Cochrane systematic reviews was found to be in need of improvement. Journal of Clinical Epidemiology, 67, 645-653.

IMBENS, G. & RUBIN, D. B. 2015. Causal inference for statistics, social, and biomedical sciences: an introduction, Cambridge, Cambridge University Press.

IMBENS, G. W. 2010. Better late than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature, 48, 399-423.

IMBENS, G. W. & LEMIEUX, T. 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics, 142, 615-635.

INGBER, S. Z., BUSER, M. C., POHL, H. R., ABADIN, H. G., EDWARD MURRAY, H. & SCINICARIELLO, F. 2013. DDT/DDE and breast cancer: a meta-analysis. Regulatory Toxicology and Pharmacology, 67, 421-433.

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER. 2012. Cancer fact sheets: breast cancer data [Online]. Available: http://gco.iarc.fr/today/fact-sheets-cancers?cancer=15&type=0&sex=2 [Accessed 7 September 2018].

INTERNATIONAL PROGRAMME ON CHEMICAL SAFETY. 2002. Global assessment on the state of the science of endocrine disruptors. Available: http://www.who.int/iris/handle/10665/67357 [Accessed 7 September 2018].

IOM (INSTITUTE OF MEDICINE) 2012. Breast cancer and the environment: a life course approach, Washington, DC, The National Academies Press.

JAKOBSSON, N., PERSSON, M. & SVENSSON, M. 2013. Class-size effects on adolescents' mental health and well-being in Swedish schools. Education Economics, 21, 248-263.

JANSSENS, W. 2011. Externalities in program evaluation: the impact of a women's empowerment program on immunization. Journal of the European Economic Association, 9, 1082-1113.

JENICEK, M. 1997. Epidemiology, evidence-based medicine, and evidence-based public health. Journal of Epidemiology, 7, 187-197.

JENKINS, J. M. 2018. Healthy and ready to learn: effects of a school-based public health insurance outreach program for kindergarten-aged children. Journal of School Health, 88, 44-53.

JENSEN, V. M. & WUST, M. 2015. Can Caesarean section improve child and maternal health? The case of breech babies. Journal of Health Economics, 39, 289-302.

JOHANSSON, P. & PALME, M. 2005. Moral hazard and sickness insurance. Journal of Public Economics, 89, 1879-1890.

JOHN, A., HAWTON, K., OKOLIE, C., DENNIS, M., PRICE, S. F. & LLOYD, K. 2018. Means restriction for the prevention of suicide: generic protocol. Cochrane Database of Systematic Reviews.

JOHNSTON, D. W. & LEE, W. S. 2009. Retiring to the good life? The short-term effects of retirement on health. Economics Letters, 103, 8-11.

JOHNSTON, D. W., LORDON, G., SHIELDS, M. A. & SUZIEDELYTE, A. 2015. Education and health knowledge: evidence from UK compulsory schooling reform. Social Science and Medicine, 127, 92-100.

JUDGE, S. 2011. The effect of parental involvement laws on the timing of teenagers' abortions. Contraception, 84 (3), 307.

JUDGE, S. 2012. The effect of parental involvement laws on the likelihood of later-term abortion procedures for minors. Contraception, 86 (3), 293.

JURGES, H., KRUK, E. & REINHOLD, S. 2010. The Effect of Compulsory Schooling on Health--Evidence from Biomarkers. CESifo Group Munich, CESifo Working Paper Series: CESifo Working Paper No. 3105.

KADIYALA, S. & STRUMPF, E. 2011. How Effective is Population-Based Cancer Screening? Regression Discontinuity Estimates from the US Guideline Screening Initiation Ages. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1893793 [Accessed 7 September 2018].

KADIYALA, S. & STRUMPF, E. C. 2009. Guidelines and cancer screening in the United States and Canadian health systems. Value in Health, 12 (3), A56.

KADIYALA, S. & STRUMPF, E. C. 2011. Are United States and Canadian cancer screening rates consistent with guideline information regarding the age of screening initiation? International Journal for Quality in Health Care, 23, 611-620.

KATIKIREDDI, S. V., BOND, L. & HILTON, S. 2014. Perspectives on econometric modelling to inform policy: a UK qualitative case study of minimum unit pricing of alcohol. European Journal of Public Health, 24, 490-5.

KATIKIREDDI, S. V., HIGGINS, M., BOND, L., BONELL, C. & MACINTYRE, S. 2011. How evidence based is English public health policy? BMJ, 343, d7310.

KELLY, M. P., MORGAN, A., BONNEFOY, J., BUTT, J. & BERGMAN, V. 2007. The social determinants of health: Developing an evidence base for political action. Final Report to World Health Organization Commission on the Social Determinants of Health Universidad del Desarrollo, Chile, and National Institute for Health and Clinical Excellence, UK.

KELLY, S. E., MOHER, D. & CLIFFORD, T. J. 2016. Quality of conduct and reporting in rapid reviews: an exploration of compliance with PRISMA and AMSTAR guidelines. Systematic Reviews, 5, 79.

KHANJANI, N., HOVING, J. L., FORBES, A. B. & SIM, M. R. 2007. Systematic review and meta-analysis of cyclodiene insecticides and breast cancer. Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews, 25, 23-52.

KLINGEMANN, H. 2001. Alcohol and its social consequences – the forgotten dimension [Online]. World Health Organization Regional Office for Europe. Available: http://www.unicri.it/min.san.bollettino/dati/E76235.pdf [Accessed June 23 2018].

KNAI, C., PETTICREW, M., DURAND, M. A., EASTMURE, E. & MAYS, N. 2015. Are the Public Health Responsibility Deal alcohol pledges likely to improve public health? An evidence synthesis. Addiction, 110, 1232-46.

KOCH, S. F. 2012. The Abolition of User Fees and the Demand for Health Care: Re-evaluating the Impact. Economic Research Southern Africa.

KOCH, S. F. 2013. User Fee Abolition in South Africa: Re-Evaluating the Impact? : University of Pretoria, Department of Economics.

KOCH, S. F. & RACINE, J. S. 2013. Health Care Facility Choice and User Fee Abolition: Regression Discontinuity in a Multinomial Choice Setting. McMaster University, Department of Economics Working Papers.

KOCH, T. G. 2013. Using RD design to understand heterogeneity in health insurance crowd-out. Journal of Health Economics, 32, 599-611.

KOHATSU, N. D., ROBINSON, J. G. & TORNER, J. C. 2004. Evidence-based public health - An evolving concept. American Journal of Preventive Medicine, 27, 417-421.

KONG, A. 2011. Three Essays in Health Economics. Ph.D., Simon Fraser University.

KOPPA, V. 2018. The effect of alcohol access on sexually transmitted diseases: Evidence from the minimum legal drinking age. American Journal of Health Economics, 4, 164-184.

KUSS, O., LEGLER, T. & BORGERMANN, J. 2011. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. Journal of Clinical Epidemiology, 64, 1076-84.

KYPRI, K., DAVIE, G., MCELDUFF, P., LANGLEY, J. & CONNOR, J. 2017. Long-term effects of lowering the alcohol minimum purchasing age on traffic crash injury rates in New Zealand. Drug and Alcohol Review, 36, 178-185.

KYPRI, K., VOAS, R. B., LANGLEY, J. D., STEPHENSON, S. C., BEGG, D. J., TIPPETTS, A. S. & DAVIE, G. S. 2006. Minimum purchasing age for alcohol and traffic crash injuries among 15- to 19-year-olds in New Zealand. American Journal of Public Health, 96, 126-31.

LABRECQUE, J. A. & KAUFMAN, J. S. 2016. Commentary: Can a quasi-experimental design be a better idea than an experimental one? Epidemiology, 27, 500-502.

LACHAT, C., HAWWASH, D., OCKÉ, M. C., BERG, C., FORSUM, E., HÖRNELL, A., LARSSON, C., SONESTEDT, E., WIRFÄLT, E., ÅKESSON, A., KOLSTEREN, P., BYRNES, G., DE KEYZER, W., VAN CAMP, J., CADE, J. E., SLIMANI, N., CEVALLOS, M., EGGER, M. & HUYBRECHTS, I. 2016. Strengthening the Reporting of Observational Studies in Epidemiology—Nutritional Epidemiology (STROBE-nut): an extension of the STROBE Statement. PLoS Medicine, 13, e1002036.

LAMADRID-FIGUEROA, H., ANGELES, G., MROZ, T., URQUIETA-SALOMON, J., HERNANDEZ-PRADO, B., CRUZ-VALDEZ, A. & TELLEZ-ROJO, M. M. 2008. Impact of Oportunidades on contraceptive methods use in adolescent and young adult women living in rural areas, 1997-2000. MEASURE Evaluation Working Paper Series. Chapel Hill, North Carolina: University of North Carolina at Chapel Hill, Carolina Population Center.

LAMMERS, M., BLOEMEN, H. & HOCHGUERTEL, S. 2013. Job search requirements for older unemployed: Transitions to employment, early retirement and disability benefits. European Economic Review, 58, 31-57.

LANZA, S. T., MOORE, J. E. & BUTERA, N. M. 2013. Drawing causal inferences using propensity scores: a practical guide for community psychologists. American Journal of Community Psychology, 52, 380-392.

LAUBY-SECRETAN, B., LOOMIS, D., GROSSE, Y., GHISSASSI, F. E., BOUVARD, V., BENBRAHIM-TALLAA, L., GUHA, N., BAAN, R., MATTOCK, H. & STRAIF, K. 2013. Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. The Lancet Oncology, 14, 287-288.

LAVIS, J. N. 2009. How can we support the use of systematic reviews in policymaking? PLoS Medicine, 6, e1000141.

LEE, D. S. & LEMIEUX, T. 2010. Regression discontinuity designs in economics. Journal of Economic Literature, 48, 281-355.

LENG, L., LI, J., LUO, X. M., KIM, J. Y., LI, Y. M., GUO, X. M., CHEN, X., YANG, Q. Y., LI, G. & TANG, N. J. 2016. Polychlorinated biphenyls and breast cancer: A congener-specific meta-analysis. Environment International, 88, 133-141.

LI, M. C., CHEN, P. C., TSAI, P. C., FURUE, M., ONOZUKA, D., HAGIHARA, A., UCHI, H., YOSHIMURA, T. & GUO, Y. L. 2015. Mortality after exposure to polychlorinated biphenyls and polychlorinated dibenzofurans: A meta-analysis of two highly exposed cohorts. International Journal of Cancer, 137, 1427-1432.

LINDEBOOM, M., LLENA-NOZAL, A. & VAN DER KLAAUW, B. 2009. Parental education and child health: Evidence from a schooling reform. Journal of Health Economics, 28, 109-131.

LINDEN, A. & ADAMS, J. L. 2011. Applying a propensity score-based weighting model to interrupted time series data: improving causal inference in programme evaluation. Journal of Evaluation in Clinical Practice, 17, 1231-1238.

LINDEN, A. & ADAMS, J. L. 2012. Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. Journal of Evaluation in Clinical Practice, 18, 317-325.

LINDO, J. M., SIMINSKI, P. & YEROKHIN, O. 2014. Breaking the Link Between Legal Access to Alcohol and Motor Vehicle Accidents: Evidence from New South Wales. National Bureau of Economic Research, Inc, NBER Working Papers: 19857.

LIPSEY, M. W., FARRAN, D. C., BILBREY, C., HOFER, K. G. & DONG, N. 2011. Initial Results of the Evaluation of the Tennessee Voluntary PreK Program [Online]. Peabody Research Institute, Vanderbilt University. Available: https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/April2011_PRI_Initial_TN-VPK_ProjectResults.pdf [Accessed 27 August 2015].

LITTLE, J., HIGGINS, J. P. T., IOANNIDIS, J. P. A., MOHER, D., GAGNON, F., VON ELM, E., KHOURY, M. J., COHEN, B., DAVEY-SMITH, G., GRIMSHAW, J., SCHEET, P., GWINN, M., WILLIAMSON, R. E., ZOU, G. Y., HUTCHINGS, K., JOHNSON, C. Y., TAIT, V., WIENS, M., GOLDING, J., VAN DUIJN, C., MCLAUGHLIN, J., PATERSON, A., WELLS, G., FORTIER, I., FREEDMAN, M., ZECEVIC, M., KING, R., INFANTE-RIVARD, C., STEWART, A. & BIRKETT, N. 2009. Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. European Journal of Epidemiology, 24, 37-55.

LIU, X. Does the US Supplemental Nutrition Assistance Program Contribute to Adult Obesity? Evidence from Regression Discontinuity[poster]. Health & Healthcare in America: From Economics to Policy, 2014. Ashecon.

LLERAS-MUNEY, A. 2005. The relationship between education and adult mortality in the United States. The Review of Economic Studies, 72, 189-221.

LOPEZ BERNAL, J., CUMMINS, S. & GASPARRINI, A. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. International Journal of Epidemiology, 46, 348-355.

LORENC, T., TYNER, E. F., PETTICREW, M., DUFFY, S., MARTINEAU, F. P., PHILLIPS, G. & LOCK, K. 2014. Cultures of evidence across policy sectors: systematic review of qualitative evidence. European Journal of Public Health, 24, 1041-1047.

LU, X. 2013. The Effects of Car Driving and Purchasing Restrictions on Air Quality and the Use of Public Transportation in Beijing, China. M.Sc., Tufts University.

LUDWIG, J. & MILLER, D. L. 2007. Does head start improve children's life chances? Evidence from a regression discontinuity design. Quarterly Journal of Economics, 122, 159-208.

LUNNY, C., BRENNAN, S. E., MCDONALD, S. & MCKENZIE, J. E. 2017. Toward a comprehensive evidence map of overview of systematic review methods: paper 1—purpose, eligibility, search and data extraction. Systematic Reviews, 6, 1-27.

MAAS, I. L., NOLTE, S., WALTER, O. B., BERGER, T., HAUTZINGER, M., HOHAGEN, F., LUTZ, W., MEYER, B., SCHRODER, J., SPATH, C., KLEIN, J. P., MORITZ, S. & ROSE, M. 2017. The regression discontinuity design showed to be a valid alternative to a randomized controlled trial for estimating treatment effects. Journal of Clinical Epidemiology, 82, 94-102.

MACINTYRE, S. 2011. Good intentions and received wisdom are not good enough: the need for controlled trials in public health. Journal of Epidemiology and Community Health, 65, 564-7.

MACMILLAN CANCER SUPPORT. 2013. Receptors for breast cancer [Online]. Available: https://www.macmillan.org.uk/information-and-support/breast-cancer/treating/treatment-decisions/understanding-your-diagnosis/receptors-for-breast-cancer.html [Accessed September 7 2018].

MARCUS, S. M., WEAVER, J., LIM, S., DUAN, N., GIBBONS, R. D. & ROSENHECK, R. 2012. Assessing the causal effect of Section 8 housing vouchers as the active ingredient for decreasing homelessness in veterans with mental illness. Health Services and Outcomes Research Methodology, 12, 273-287.

MARIER, A. 2014. Where does the money go? analyzing the patient experience in safety-net hospitals. Value in Health, 17, 231-237.

MARK, B. A., HARLESS, D. W., SPETZ, J., REITER, K. L. & PINK, G. H. 2013. California's minimum nurse staffing legislation: results from a natural experiment. Health Services Research, 48, 435-454.

MARTENS, E. P., PESTMAN, W. R., DE BOER, A., BELITSER, S. V. & KLUNGEL, O. H. 2006. Instrumental variables: application and limitations. Epidemiology, 17, 260-267.

MARTIN, A., OGILVIE, D. & SUHRCKE, M. 2014. Evaluating causal relationships between urban built environment characteristics and obesity: a methodological review of observational studies. International Journal of Behavioral Nutrition & Physical Activity, 11, 142.

MARTINEAU, F., TYNER, E., LORENC, T., PETTICREW, M. & LOCK, K. 2013. Population-level interventions to reduce alcohol-related harm: An overview of systematic reviews. Preventive Medicine, 57, 278-296.

MCCARTT, A. T., HELLINGA, L. A. & KIRLEY, B. B. 2010. The effects of minimum legal drinking age 21 laws on alcohol-related driving in the United States. Journal of Safety Research, 41, 173-81.

MCCRARY, J. & ROYER, H. 2003. Does Maternal Education Affect Infant Health? A Regression Discontinuity Approach Based on School Age Entry Laws.

MCCRARY, J. & ROYER, H. 2011. The Effect of Female Education on Fertility and Infant Health: Evidence from School Entry Policies Using Exact Date of Birth. American Economic Review, 101, 158-195.

MCFARLANE, W. R., LEVIN, B., TRAVIS, L., LUCAS, F. L., LYNCH, S., VERDI, M., WILLIAMS, D., ADELSHEIM, S., CALKINS, R., CARTER, C. S., CORNBLATT, B., TAYLOR, S. F., AUTHER, A. M., MCFARLAND, B., MELTON, R., MIGLIORATI, M., NIENDAM, T., RAGLAND, J. D., SALE, T., SALVADOR, M. & SPRING, E. 2015. Clinical and functional outcomes after 2 years in the early detection and intervention for the prevention of psychosis multisite effectiveness trial. Schizophrenia Bulletin, 41, 30-43.

MCKENZIE, J. E. & BRENNAN, S. E. 2017. Overviews of systematic reviews: great promise, greater challenge. Systematic Reviews, 6, 185.

MEDINA, C., NUNEZ, J. & TAMAYO, J. A. 2013. The Unemployment Subsidy Program in Colombia: An Assessment. Inter-American Development Bank, Research Department.

MELLER, M. & LITSCHIG, S. 2014. Saving lives: Evidence from a conditional food supplementation program. Journal of Human Resources, 49, 1014-1052.

MEYER, B. D. 1995. Natural and quasi-experiments in economics. Journal of Business & Economic Statistics, 13, 151-161.

MEZUK, B., LARKIN, G. L., PRESCOTT, M. R., TRACY, M., VLAHOV, D., TARDIFF, K. & GALEA, S. 2009. The influence of a major disaster on suicide risk in the population. Journal of Traumatic Stress, 22, 481-488.

MILLER, G., PINTO, D. & VERA-HERNÁNDEZ, M. 2013. Risk protection, service use, and health outcomes under Colombia's health insurance program for the poor. American Economic Journal: Applied Economics, 5, 61-91.

MILLER, G., PINTO, D. M. & VERA-HERNANDEZ, M. 2009. High-Powered Incentives in Developing Country Health Insurance: Evidence from Colombia's Regimen Subsidiado. National Bureau of Economic Research, Inc, NBER Working Papers: 15456.

MIRON JEFFREY, A. & TETELBAUM, E. 2009. Does the minimum legal drinking age save lives? Economic Inquiry, 47, 317-336.

MITOMA, C., UCHI, H., TSUKIMORI, K., YAMADA, H., AKAHANE, M., IMAMURA, T., UTANI, A. & FURUE, M. 2015. Yusho and its latest findings-A review in studies conducted by the Yusho Group. Environment International, 82, 41-48.

MOBERG, J., OXMAN, A. D., ROSENBAUM, S., SCHÜNEMANN, H. J., GUYATT, G., FLOTTORP, S., GLENTON, C., LEWIN, S., MORELLI, A., RADA, G., ALONSO-COELLO, P., MOBERG, J., OXMAN, A., COELLO, P. A., SCHÜNEMANN, H., GUYATT, G., ROSENBAUM, S., MORELLI, A., AKL, E., GLENTON, C., GULMEZOGLU, M., FLOTTORP, S., LEWIN, S., MUSTAFA, R. A., RADA, G., SINGH, J., VON ELM, E., VOGEL, J., WATINE, J. & FOR THE, G. W. G. 2018. The GRADE Evidence to Decision (EtD) framework for health system and public health decisions. Health Research Policy and Systems, 16, 45.

MOHER, D., JADAD, A. R., NICHOL, G., PENMAN, M., TUGWELL, P. & WALSH, S. 1995. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled Clinical Trials, 16, 62-73.

MOHER, D., SHAMSEER, L., CLARKE, M., GHERSI, D., LIBERATI, A., PETTICREW, M., SHEKELLE, P. & STEWART, L. A. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Review, 4, 1.

MONSTAD, K., PROPPER, C. & SALVANES, K. G. 2008. Education and fertility: evidence from a natural experiment. Scandinavian Journal of Economics, 110, 827-852.

MOORE, L. & MOORE, G. F. 2011. Public health evaluation: which designs work, for whom and under what circumstances? Journal of Epidemiology and Community Health, 65, 596-7.

MORGAN, O. W., GRIFFITHS, C. & MAJEED, A. 2007. Interrupted time-series analysis of regulations to reduce paracetamol (acetaminophen) poisoning. PLoS Medicine, 4, e105.

MOSCOE, E., BOR, J. & BÄRNIGHAUSEN, T. 2015. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. Journal of Clinical Epidemiology, 68, 132-143.

MOULY, T. A. & TOMS, L. L. 2016. Breast cancer and persistent organic pollutants (excluding DDT): a systematic literature review. Environmental Science & Pollution Research, 23, 22385-22407.

MUHLESTEIN, D. B. & SEIBER, E. E. 2013. State variability in children's medicaid/chip crowd-out estimates. Medicare and Medicaid Research Review, 3, E1-E22.

NABERNEGG, M. 2012. El impacto del BDH en el gasto de bienes no deseados: Un análisis de regresión discontinua. University Library of Munich, Germany.

NAKAMURA, R. 2011. Essays on the economics of obesity. Ph.D., University of York (United Kingdom).

NAKAMURA, R. 2012. Intergenerational effect of schooling and childhood overweight. HEDG, c/o Department of Economics, University of York.

NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE (NICE). 2012. Developing NICE guidelines: the manual. PMG20 [Online]. Available: https://www.nice.org.uk/process/pmg20/chapter/introduction-and-overview [Accessed September 13 2018].

NAZIF-MUNOZ, J. I., FALCONER, J. & GONG, A. 2017. Are child passenger fatalities and child passenger severe injuries equally affected by child restraint legislation? The case of Chile. International Journal of Injury Control & Safety Promotion, 24, 501-509.

NEIDELL, M. 2009. Information, avoidance behavior, and health: The effect of ozone on asthma hospitalizations. Journal of Human Resources, 44, 450-478.

NEIDELL, M. 2010. Air quality warnings and outdoor activities: Evidence from Southern California using a regression discontinuity design. Journal of Epidemiology and Community Health, 64, 921-926.

NELSON, J. P. & MCNALL, A. D. 2016. Alcohol prices, taxes, and alcohol-related harms: A critical review of natural experiments in alcohol policy for nine countries. Health Policy (Amsterdam, Netherlands), 120, 264-72.

NELSON, J. P. & MCNALL, A. D. 2017. What happens to drinking when alcohol policy changes? A review of five natural experiments for alcohol taxes, prices, and availability. European Journal of Health Economics, 18, 417-434.

NHS HEALTH SCOTLAND. 2018. Monitoring and evaluating Scotland's Alcohol Strategy (MESAS). Monitoring report 2018 [Online]. Available: http://www.healthscotland.scot/media/1863/mesas-monitoring-report-2018.pdf [Accessed June 20 2018].

NIKOLOVA, S. 2010. Health insurance transitions of SCHIP-eligible children in response to higher public premiums. Ph.D., University of North Carolina at Chapel Hill.

NIKOLOVA, S. & STEARNS, S. 2014. The impact of CHIP premium increases on insurance outcomes among CHIP eligible children. BMC Health Services Research, 14, 101.

NISHI, A., MICHAEL MCWILLIAMS, J., NOGUCHI, H., HASHIMOTO, H., TAMIYA, N. & KAWACHI, I. 2012. Health benefits of reduced patient cost sharing in Japan. Bulletin of the World Health Organization, 90, 426-435A.

NOONAN, D. S. 2014. Smoggy with a chance of altruism: The effects of ozone alerts on outdoor recreation and driving in Atlanta. Policy Studies Journal, 42, 122-145.

OGILVIE, D., HAMILTON, V., EGAN, M. & PETTICREW, M. 2005. Systematic reviews of health effects of social interventions: 1. Finding the evidence: how far should you go? Journal of Epidemiology and Community Health, 59, 804-808.

O'KEEFFE, A. G., GENELETTI, S., BAIO, G., SHARPLES, L. D., NAZARETH, I. & PETERSEN, I. 2014. Regression discontinuity designs: an approach to the evaluation of treatment efficacy in primary care using observational data. BMJ, 349, g5293.

OLDENBURG, C. E., MOSCOE, E. & BARNIGHAUSEN, T. 2016. Regression discontinuity for causal effect estimation in epidemiology. Current Epidemiology Reports, 3, 233-241.

OLSHO, L. E. W., KLERMAN, J. A., RITCHIE, L., WAKIMOTO, P., WEBB, K. L. & BARTLETT, S. 2015. Increasing child fruit and vegetable intake: findings from the US Department of Agriculture Fresh Fruit and Vegetable Program. Journal of the Academy of Nutrition and Dietetics, 115, 1283-1290.

PAGE, M. J. & MOHER, D. 2017. Evaluations of the uptake and impact of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement and extensions: a scoping review. Systematic Reviews, 6, 263.

PALANGKARAYA, A. & YONG, J. 2007. How effective is "lifetime health cover" in raising private health insurance coverage in Australia? An assessment using regression discontinuity. Applied Economics, 39, 1361-1374.

PALMER, M., MITRA, S., MONT, D. & GROCE, N. 2015. The impact of health insurance for children under age 6 in Vietnam: A regression discontinuity approach. Social Science and Medicine, 145, 217-226.

PARK, J. H., CHA, E. S., KO, Y., HWANG, M. S., HONG, J. H. & LEE, W. J. 2014. Exposure to dichlorodiphenyltrichloroethane and the risk of breast cancer: a systematic review and meta-analysis. Osong Public Health & Research Perspectives, 5, 77-84.

PARK, W. 2013. Essays on the Returns to Higher Education. Ph.D., Columbia University.

PATRICK, H. & KLEIN, T. J. 2010. The effect of private health insurance on medical care utilization and self-assessed health in germany. Health Economics, 19, 1048-1062.

PECKHAM, J. G. & KROPP, J. D. 2012. Are National School Lunch Program Participants More Likely to be Obese? Dealing with Identification. Agricultural and Applied Economics Association.

PEIKES, D. N., MORENO, L. & ORZOL, S. M. 2008. Propensity score matching: A note of caution for evaluators of social programs. American Statistician, 62, 222-231.

PESKO, M. 2012. The Effects of National Disasters on Stress and Substance Use in the United States. Ph.D., University of Illinois, Chicago.

PESKO, M. F. 2014. Stress and smoking: Associations with terrorism and causal impact. Contemporary Economic Policy, 32, 351-371.

PETTICREW, M. 2013. Public health evaluation: epistemological challenges to evidence production and use. Evidence and Policy, 9, 87-95.

PETTICREW, M., CUMMINS, S., FERRELL, C., FINDLAY, A., HIGGINS, C., HOY, C., KEARNS, A. & SPARKS, L. 2005. Natural experiments: an underused tool for public health? Public Health, 119, 751-757.

PETTICREW, M., MAANI HESSARI, N., KNAI, C. & WEIDERPASS, E. 2017. How alcohol industry organisations mislead the public about alcohol and cancer. Drug and Alcohol Review, 37, 293-303.

PETTICREW, M. & ROBERTS, H. 2003. Evidence, hierarchies, and typologies: horses for courses. Journal of Epidemiology and Community Health, 57, 527-529.

PETTICREW, M. & ROBERTS, H. 2006. Systematic reviews in the social sciences: a practical guide, Oxford, Blackwell.

PIERCE, L., DAHL, M. S. & NIELSEN, J. 2013. In sickness and in wealth: psychological and sexual costs of income comparison in marriage. Personality and Social Psychology Bulletin, 39, 359-374.

PIERONI, L., CHIAVARINI, M., MINELLI, L. & SALMASI, L. 2012. The role of smoking bans on cigarettes and alcohol habits in Italy. European Journal of Epidemiology, 27, S28.

PIERONI, L., CHIAVARINI, M., MINELLI, L. & SALMASI, L. 2013. The role of anti-smoking legislation on cigarette and alcohol consumption habits in Italy. Health Policy, 111, 116-126.

PIERONI, L. & SALMASI, L. 2012. Does cigarette smoking affect body weight? causal estimates from the clean indoor air law discontinuity. University Library of Munich, Germany.

PIERONI, L. & SALMASI, L. 2015. Does cigarette smoking affect body weight? causal estimates from the clean indoor air law discontinuity. Economica, 82, 671-704.

PITT, M. M., KHANDKER, S. R., MCKERNAN, S.-M. & LATIF, M. A. 1999. Credit Programs for the Poor and Reproductive Behavior in Low-Income Countries: Are the Reported Causal Relationships the Result of Heterogeneity Bias? Demography, 36, 1-21.

POPAY, J., ROBERTS, H., SOWDEN, A., PETTICREW, M., ARAI, L., RODGERS, M. & BRITTEN, N. 2006. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC Methods Programme.

PORTA, M. 2014. A Dictionary of epidemiology, Oxford, Oxford University Press.

POWDTHAVEE, N. 2010. Does education reduce the risk of hypertension? Estimating the biomarker effect of compulsory schooling in England. Journal of Human Capital, 4, 173-202.

PROPPER, C., SUTTON, M., WHITNALL, C. & WINDMEIJER, F. 2010. Incentives and targets in hospital care: evidence from a natural experiment. Journal of Public Economics, 94, 318-335.

PUBLIC HEALTH ENGLAND. 2016. The public health burden of alcohol and the effectiveness and cost-effectiveness of alcohol control policies: annexes [Online]. London. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/574054/alcohol_public_health_burden_annexes.pdf [Accessed August 14 2018].

RAHMAN, M. M. 2014. Estimating the average treatment effect of social safety net programmes in Bangladesh. Journal of Development Studies, 50, 1550-1569.

RAJBHAR, M. & MOHANTY, S. K. 2017. Reproductive and child health services and demographic change in the districts of Uttar Pradesh, 2002-13. Journal of Biosocial Science, 49, 685-709.

RAMSAY, C. R., MATOWE, L., GRILLI, R., GRIMSHAW, J. M. & THOMAS, R. E. 2003. Interrupted time series designs in health technology assessment: Lessons from two systematic reviews of behavior change strategies. International Journal of Technology Assessment in Health Care, 19, 613-623.

RASHAD, H. 1992. The mortality impact of oral rehydration therapy in Egypt: re-appraisal of evidence. Baltimore, Maryland, Johns Hopkins University, School of Hygiene and Public Health, Institute for International Programs, 1992 Oct.

RASSEN, J. A., BROOKHART, M. A., GLYNN, R. J., MITTLEMAN, M. A. & SCHNEEWEISS, S. 2009. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. Journal of Clinical Epidemiology, 62, 1226-1232.

REHFUESS, E. A. & AKL, E. A. 2013. Current experience with applying the GRADE approach to public health interventions: an empirical study. BMC Public Health, 13, 9.

REHM, J., MATHERS, C., POPOVA, S., THAVORNCHAROENSAP, M., TEERAWATTANANON, Y. & PATRA, J. 2009. Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. The Lancet, 373, 2223-2233.

RIECK, K. M. E. 2012. DOES CHILD CARE AFFECT PARENTS' SICKNESS ABSENCE? EVIDENCE FROM A NORWEGIAN PATERNITY LEAVE REFORM. University of Bergen, Department of Economics.

ROBERTS, C. & TORGERSON, D. 1998. Randomisation methods in controlled trials. BMJ, 317, 1301.

ROCKERS, P. C., ROTTINGEN, J. A., SHEMILT, I., TUGWELL, P. & BARNIGHAUSEN, T. 2015. Inclusion of quasi-experimental studies in systematic reviews of health systems research. Health Policy, 119, 511-21.

RODGERS, K. M., UDESKY, J. O., RUDEL, R. A. & BRODY, J. G. 2018. Environmental chemicals and breast cancer: an updated review of epidemiological literature informed by biological mechanisms. Environmental Research, 160, 152-182.

ROSE, G. 1981. Strategy of prevention: lessons from cardiovascular disease. BMJ, 282, 1847.

ROSE, G. 2001. Sick individuals and sick populations. International Journal of Epidemiology, 30, 427-432.

ROSENBAUM, P. R. & RUBIN, D. B. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

ROSERO, J. & OOSTERBEEK, H. 2011. Trade-offs between Different Early Childhood Interventions: Evidence from Ecuador. Tinbergen Institute, Tinbergen Institute Discussion Papers: 11-102/3.

RUBIN, D. B. 1997. Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine, 127, 757-763.

RYAN, A., SUTTON, M. & DORAN, T. 2014. Does winning a pay-for-performance bonus improve subsequent quality performance? Evidence from the hospital quality incentive demonstration. Health Services Research, 49, 568-587.

SACKETT, D. L., ROSENBERG, W. M., GRAY, J. A., HAYNES, R. B. & RICHARDSON, W. S. 1996. Evidence based medicine: what it is and what it isn't. BMJ, 312, 71-2.

SACKETT, D. L. & WENNBERG, J. E. 1997. Choosing the best research design for each question. BMJ, 315, 1636.

SADASIVAIAH, S., TOZAN, Y. & BREMAN, J. G. 2007. Dichlorodiphenyltrichloroethane (DDT) for indoor residual spraying in Africa: How can it be used for malaria control? American Journal of Tropical Medicine and Hygiene, 77, 249-263.

SAMARAKOON, S. & PARINDURI, R. A. 2015. Does education empower women? evidence from Indonesia. World Development, 66, 428-442.

SANDERS, N. J. & STOECKER, C. 2015. Where have all the young men gone? Using sex ratios to measure fetal death rates. Journal of Health Economics, 41, 30-45.

SANTOS, R. G. 2006. Effectiveness of early intervention for infants and their families: Relating the working alliance to program outcomes. Ph.D., University of Manitoba.

SCHANZENBACH, D. 2005. Do School Lunches Contribute to Childhood Obesity? : Harris School of Public Policy Studies, University of Chicago, Working Papers: 0513.

SCHANZENBACH, D. 2009. Do school lunches contribute to childhood obesity? Journal of Human Resources, 44, 684-709.

SCHOCHET, P., COOK, T. D., DEKE, J., IMBENS, G. W., LOCKWOOD, J. R., PORTER, J. & SMITH, J. 2010. Standards for regression discontinuity designs. Available: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf. [Accessed 2015 September]

SCHÜNEMANN, H. J., CUELLO, C., AKL, E. A., MUSTAFA, R. A., MEERPOHL, J. J., THAYER, K., MORGAN, R. L., GARTLEHNER, G., KUNZ, R., KATIKIREDDI, S. V.,

STERNE, J., HIGGINS, J. P. T. & GUYATT, G. 2018. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. Journal of Clinical Epidemiology [in press].

SCHÜNEMANN, H. J., MUSTAFA, R., BROZEK, J., SANTESSO, N., ALONSO-COELLO, P., GUYATT, G., SCHOLTEN, R., LANGENDAM, M., LEEFLANG, M. M., AKL, E. A., SINGH, J. A., MEERPOHL, J., HULTCRANTZ, M., BOSSUYT, P., OXMAN, A. D., SCHÜNEMANN, H. J., MUSTAFA, R., BROZEK, J., SANTESSO, N., ALONSO-COELLO, P., SCHOLTEN, R., LANGENDAM, M., BOSSUYT, P., LEEFLANG, M. M., AKL, E. A., SINGH, J., MEERPOHL, J., HULTCRANTZ, M., GUYATT, G., OXMAN, A. D., LANGE, S., PARMELLI, E., MOBERG, J., ROSENBAUM, S., BRIGNARDELLO-PETERSEN, R., WIERCIOCH, W., DAVOLI, M., NOWAK, A. & DIETL, B. 2016. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. Journal of Clinical Epidemiology, 76, 89-98.

SCOTTISH GOVERNMENT. 2016. Reducing the damaging impact of drugs and alcohol [Online]. Available: http://www.gov.scot/Topics/Justice/policies/drugs-alcohol [Accessed June 20 2018].

SCOTTISH GOVERNMENT. 2018. Alcohol [Online]. Available: http://www.gov.scot/Topics/Health/Services/Alcohol [Accessed June 20 2018].

SCOTTISH HEALTH ACTION ON ALCOHOL PROBLEMS. 2014. Alcohol and the developing adolescent brain: evidence review [Online]. Available: http://www.shaap.org.uk/images/shaap_developing_adolescents_brain_press.pdf [Accessed June 30 2018].

SCOTTISH INTERCOLLEGIATE GUIDELINES NETWORK. 2015. SIGN 50: a guideline developer's handbook [Online]. Available: https://www.sign.ac.uk/sign-50.html [Accessed September 13 2018].

SCOTTISH PUBLIC HEALTH OBSERVATORY. 2018. Breast cancer: Scottish data [Online]. Available: https://www.scotpho.org.uk/health-wellbeing-and-disease/cancer-breast/data/scottish/ [Accessed 28 August 2018].

SHADISH, W. R., COOK, T. & CAMPBELL, D. T. 2002. Experimental and quasi-experimental designs for generalized causal inference, Boston, Houghton Mifflin.

SHADISH, W. R. & STEINER, P. M. 2010. A primer on propensity score analysis. Newborn and Infant Nursing Reviews, 10, 19-26.

SHAW, S. D., BLUM, A., WEBER, R., KANNAN, K., RICH, D., LUCAS, D., KOSHLAND, C. P., DOBRACA, D., HANSON, S. & BIRNBAUM, L. S. 2010. Halogenated flame

retardants: Do the fire safety benefits justify the risks? Reviews on Environmental Health, 25, 261-305.

SHEA, B. J., REEVES, B. C., WELLS, G., THUKU, M., HAMEL, C., MORAN, J., MOHER, D., TUGWELL, P., WELCH, V., KRISTJANSSON, E. & HENRY, D. A. 2017. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, 358, j4008.

SHIGEOKA, H. 2014. The effect of patient cost sharing on utilization, health, and risk protection. American Economic Review, 104, 2152-2184.

SHULTS, R. A., ELDER, R. W., SLEET, D. A., NICHOLS, J. L., ALAO, M. O., CARANDE-KULIS, V. G., ZAZA, S., SOSIN, D. M., THOMPSON, R. S., TASK FORCE COMMUNITY PREVENTIVE, S. & TASK FORCE ON COMMUNITY PREVENTIVE, S. 2001. Reviews of evidence regarding interventions to reduce alcohol-impaired driving. American Journal of Preventive Medicine, 21, 66-88.

SIAPLAY, M. 2012. The Impact of Social Cash Transfers on Young Adults' Labor Force Participation, Schooling, and Sexual Behaviors in South Africa. Ph.D., Oklahoma State University.

SIEGFRIED, N., PIENAAR, D. C., ATAGUBA, J. E., VOLMINK, J., KREDO, T., JERE, M. & PARRY, C. D. H. 2014. Restricting or banning alcohol advertising to reduce alcohol consumption in adults and adolescents. Cochrane Database of Systematic Reviews.

SIERING, U., EIKERMANN, M., HAUSNER, E., HOFFMANN-EßER, W. & NEUGEBAUER, E. A. 2013. Appraisal tools for clinical practice guidelines: a systematic review. PLOS ONE, 8, e82915.

SILLES, M. A. 2009. The causal effect of education on health: Evidence from the United Kingdom. Economics of Education Review, 28, 122-128.

SLOAN, F. A. & HANRAHAN, B. W. 2014. The effects of technological advances on outcomes for elderly persons with exudative age-related macular degeneration. JAMA Ophthalmology, 132, 456-463.

SMITH, K. 2013. Beyond evidence-based policy in public health: the interplay of ideas, Basingstoke, Palgrave Macmillan.

SMITH, L. 2013. The regression discontinuity design: a novel approach to assessing the real-world effectiveness of human papillomavirus (HPV) vaccination on anogenital warts. Journal of Epidemiology and Community Health, 67, e2.

SMITH, L. M., KAUFMAN, J. S., STRUMPF, E. C. & LEVESQUE, L. E. 2015. Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual

behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study. CMAJ Canadian Medical Association Journal, 187, E74-81.

SMITH, L. M., STRUMPF, E. C., KAUFMAN, J. S. & LEVESQUE, L. E. 2013. A novel approach to assessing the real-world effectiveness of the human papillomavirus vaccine: The regression discontinuity design. Pharmacoepidemiology and Drug Safety, 22, 445-446.

SNYDER, S. E. & EVANS, W. N. 2006. The effect of income on mortality: Evidence from the social security Notch. Review of Economics and Statistics, 88, 482-495.

SOJOURNER, A. J., TOWN, R. J., GRABOWSKI, D. C. & CHEN, M. M. 2012. Impacts of Unionization on Employment, Product Quality and Productivity: Regression Discontinuity Evidence From Nursing Homes. National Bureau of Economic Research, Inc, NBER Working Papers: 17733.

SOOD, N., BENDAVID, E., MUKHERJI, A., WAGNER, Z., NAGPAL, S. & MULLEN, P. 2014. Government health insurance for people below poverty line in India: quasi-experimental evaluation of insurance and health outcomes. BMJ, 349, g5114.

SOTOMAYOR, O. 2013. Fetal and infant origins of diabetes and ill health: Evidence from Puerto Rico's 1928 and 1932 hurricanes. Economics & Human Biology, 11, 281-293.

SPOTH, R., GREENBERG, M. & TURRISI, R. 2008. Preventive interventions addressing underage drinking: state of the evidence and steps toward public health impact. Pediatrics, 121 Suppl 4, S311-36.

STERNE, J. A. C., HERNAN, M. A., REEVES, B. C., SAVOVIC, J., BERKMAN, N. D., VISWANATHAN, M., HENRY, D., ALTMAN, D. G., ANSARI, M. T., BOUTRON, I., CARPENTER, J. R., CHAN, A. W., CHURCHILL, R., DEEKS, J. J., HROBJARTSSON, A., KIRKHAM, J., JUNI, P., LOKE, Y. K., PIGOTT, T. D., RAMSAY, C. R., REGIDOR, D., ROTHSTEIN, H. R., SANDHU, L., SANTAGUIDA, P. L., SCHUNEMANN, H. J., SHEA, B., SHRIER, I., TUGWELL, P., TURNER, L., VALENTINE, J. C., WADDINGTON, H., WATERS, E., WELLS, G. A., WHITING, P. F. & HIGGINS, J. P. T. 2016. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ, 355, i4919.

STILLMAN, S., MCKENZIE, D. & GIBSON, J. 2009. Migration and mental health: evidence from a natural experiment. Journal of Health Economics, 28, 677-687.

STUART, E. A., HUSKAMP, H. A., DUCKWORTH, K., SIMMONS, J., SONG, Z., CHERNEW, M. & BARRY, C. L. 2014. Using propensity scores in difference-in-differences

models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology, 14, 166-182.

STUCKLER, D., REEVES, A., KARANIKOLOS, M. & MCKEE, M. 2014. The health effects of the global financial crisis: can we reconcile the differing views? A network analysis of literature across disciplines. Health Economics, Policy and Law [Online]. Available: http://journals.cambridge.org/article_S1744133114000255.

SUN, A. & ZHAO, Y. 2014. Divorce, Abortion and Children's Sex Ratio: The Impact of Divorce Reform in China. Institute for the Study of Labor (IZA).

TAKKOUCHE, B., ETMINAN, M. & MONTES-MARTINEZ, A. 2005. Personal use of hair dyes and risk of cancer: a meta-analysis. JAMA, 293, 2516-25.

TAMIMI, R. M., HANKINSON, S. & LAGIOU, P. 2018. Breast cancer. In: ADAMI, H.-O., HUNTER, D. J., LAGIOU, P. & MUCCI, L. (eds.) Textbook of Cancer Epidemiology. 3 ed. New York: Oxford University Press.

TAMIMI, R. M., SPIEGELMAN, D., SMITH-WARNER, S. A., WANG, M., PAZARIS, M., WILLETT, W. C., ELIASSEN, A. H. & HUNTER, D. J. 2016. Population attributable risk of modifiable and nonmodifiable breast cancer risk factors in postmenopausal breast cancer. American Journal of Epidemiology, 184, 884-893.

THE EUROPEAN SCHOOL SURVEY PROJECT ON ALCOHOL AND OTHER DRUGS. 2015. ESPAD Report 2015 [Online]. Available: http://www.espad.org/report/situation/availability-of-substances.

THE LANCET 2017. Alcohol and cancer. The Lancet, 390, 2215.

THISTLETHWAITE, D. L. & CAMPBELL, D. T. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51, 309-317.

THOMSON, H. 2013. Improving utility of evidence synthesis for healthy public policy: the three Rs (relevance, rigor, and readability [and resources]). American Journal of Public Health, 103, E17-E23.

THOMSON, H., CRAIG, P., HILTON-BOON, M., CAMPBELL, M. & KATIKIREDDI, S. V. 2018. Applying the ROBINS-I tool to natural experiments: an example from public health. Systematic Reviews, 7, 15.

THOMSON, H. J. & THOMAS, S. 2013. The effect direction plot: Visual display of non-standardised effects across multiple outcome domains. Research Synthesis Methods, 8, 95-101.

TIBONE, K. L. 2013. Did the Mexico City policy affect pregnancy outcomes in Ethiopia? the impact of U.S. policy on reproductive health and family planning programs. M.P.P., Georgetown University.

TRICHOPOULOS, D., ADAMI, H. O., EKBOM, A., HSIEH, C. C. & LAGIOU, P. 2008. Early life events and conditions and breast cancer risk: From epidemiology to etiology. International Journal of Cancer, 122, 481-485.

TROCHIM, W. M. K. & CAPPELLERI, J. C. 1992. Cutoff assignment strategies for enhancing randomized clinical trials. Controlled Clinical Trials, 13, 190-212.

URQUIETA, J., ANGELES, G. & MROZ, T. 2009. Impact of Oportunidades on skilled attendance at delivery in rural areas. Economic Development & Cultural Change, 57, 539-558.

VAN EWIJK, R., DAYSAL, M. & TRANDAFIR, M. 2013. Obstetrician versus midwife supervision in low-risk deliveries: Does it matter for newborns' health? European Journal of Epidemiology, 28, S16.

VAN KIPPERSLUIS, H., O'DONNELL, O. & VAN DOORSLAER, E. 2011. Long-run returns to education: does schooling lead to an extended old age? Journal of Human Resources, 46, 695-721.

VANDENBROUCKE, J. P. & LE CESSIE, S. 2014. Commentary: regression discontinuity design: let's give it a try to evaluate medical and public health interventions. Epidemiology (Cambridge, Mass.), 25, 738-741.

VENKATARAMANI, A. S., BOR, J. & JENA, A. B. 2016. Regression discontinuity designs in healthcare research. BMJ, 352, i1216.

VERZULLI, R., JACOBS, R. & GODDARD, M. 2018. Autonomy and performance in the public sector: the experience of English NHS hospitals. The European Journal of Health Economics : HEPAC : Health Economics in Prevention and Care, 19, 607-626.

VIRENDRAKUMAR, B., JOLLEY, E., GORDON, I., BASCARAN, C. & SCHMIDT, E. 2016. Availability of evidence on cataract in low/middle-income settings: a review of reviews using evidence gap maps approach. British Journal of Ophthalmology, 100, 1455.

VISWANATHAN, M., PATNODE, C. D., BERKMAN, N. D., BASS, E. B., CHANG, S., HARTLING, L., MURAD, M. H., TREADWELL, J. R. & KANE, R. L. 2018. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. Journal of Clinical Epidemiology, 97, 26-34.

VON ELM, E., ALTMAN, D. G., EGGER, M., POCOCK, S. J., GØTZSCHE, P. C., VANDENBROUCKE, J. P. & FOR THE, S. I. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. PLoS Medicine, 4, e296.

WADDINGTON, H., ALOE, A. M., BECKER, B. J., DJIMEU, E. W., HOMBRADOS, J. G., TUGWELL, P., WELLS, G. & REEVES, B. 2017. Quasi-experimental study designs series—paper 6: risk of bias assessment. Journal of Clinical Epidemiology, 89, 43-52.

WAGENAAR, A. C. & TOOMEY, T. L. 2002. Effects of minimum drinking age laws: review and analyses of the literature from 1960 to 2000. Journal of Studies on Alcohol. Supplement, 206-225.

WALKEY, A. J., DRAINONI, M.-L., CORDELLA, N. & BOR, J. 2018. Advancing quality improvement with regression discontinuity designs. Annals of the American Thoracic Society, 15, 523-529.

WALLER, B. J., COHEN, J. E., FERRENCE, R., BULL, S. & ADLAF, E. M. 2003. The early 1990s cigarette price decrease and trends in youth smoking in Ontario. Canadian Journal of Public Health-Revue Canadienne De Sante Publique, 94, 31-35.

WANLESS, D. 2004. Securing good health for the whole population. London: HM Treasury.

WATERS, E. 2009. Evidence for public health decision-making: towards reliable synthesis. Bulletin of the World Health Organization, 87, 164.

WECHSLER, H. & NELSON, T. F. 2010. Will increasing alcohol availability by lowering the minimum legal drinking age decrease drinking and related consequences among youths? American Journal of Public Health, 100, 986-92.

WEILAND, C. & YOSHIKAWA, H. 2013. Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. Child Development, 84, 2112-2130.

WEITZEN, S., LAPANE, K. L., TOLEDANO, A. Y., HUME, A. L. & MOR, V. 2004. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiology and Drug Safety, 13, 841-53.

WILLEMSEN, M. C., SEGAAR, D. & VAN SCHAYCK, O. C. P. 2013. Population impact of reimbursement for smoking cessation: a natural experiment in The Netherlands. Addiction (Abingdon, England), 108, 602-604.

WILLIAMS, H. L. 2010. Essays on Technological Change in Health Care Markets. Ph.D., Harvard University.

WILLIAMS, R., ALEXANDER, G., ARMSTRONG, I., BAKER, A., BHALA, N., CAMPS-WALSH, G., CRAMP, M. E., DE LUSIGNAN, S., DAY, N., DHAWAN, A., DILLON, J., DRUMMOND, C., DYSON, J., FOSTER, G., GILMORE, I., HUDSON, M., KELLY, D., LANGFORD, A., MCDOUGALL, N., MEIER, P., MORIARTY, K., NEWSOME, P., O'GRADY, J., PRYKE, R., ROLFE, L., RICE, P., RUTTER, H., SHERON, N., TAYLOR, A., THOMPSON, J., THORBURN, D., VERNE, J., WASS, J. & YEOMAN, A. 2018. Disease burden and costs from excess alcohol consumption, obesity, and viral hepatitis: fourth report of the Lancet Standing Commission on Liver Disease in the UK. The Lancet, 391, 1097-1107.

WILLIAMS, S. V. 1990. Regression-discontinuity design in health evaluation. Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data, 145-149.

WILLIAMSON, E., MORLEY, R., LUCAS, A. & CARPENTER, J. 2012. Propensity scores: from naive enthusiasm to intuitive understanding. Statistical Methods in Medical Research, 21, 273-293.

WIMBUSH, E., MONTAGUE, S. & MULHERIN, T. 2012. Applications of contribution analysis to outcome planning and impact evaluation. Evaluation, 18, 310-329.

WITMAN, A. 2015. Public health insurance and disparate eligibility of spouses: The Medicare eligibility gap. Journal of Health Economics, 40, 10-25.

WITTWER, J. & GUTHMULLER, S. 2012. Means-tested complementary health insurance and healthcare utilisation in France: Evidence from a low-income population. Paris Dauphine University.

WONG, V. C., COOK, T. D., BARNETT, W. S. & JUNG, K. 2008. An effectiveness-based evaluation of five state pre-kindergarten programs. Journal of Policy Analysis and Management, 27, 122-154.

WORLD HEALTH ORGANIZATION. 2003. Chemical fact sheet: heptachlor and heptachlor epoxide [Online]. Available: http://www.who.int/water_sanitation_health/dwq/chemicals/heptachlorsum.pdf [Accessed September 13 2018].

WORLD HEALTH ORGANIZATION. 2010. Global strategy to reduce harmful use of alcohol [Online]. Available: http://www.who.int/substance_abuse/publications/global_strategy_reduce_harmful_use_alcohol/en/ [Accessed June 20 2018].

WORLD HEALTH ORGANIZATION. 2017. Guidelines on Integrated Care for Older People [Online]. Available: http://www.who.int/ageing/publications/guidelines-icope/en/ [Accessed September 13 2018].

WOUABE, E. D. & ARCAND, J.-L. 2010. Teacher training and HIV/AIDS prevention in West Africa: regression discontinuity design evidence from the Cameroon. Health Economics, 19, 36-54.

WÜST, M. & JENSEN, V. M. 2012. Essays on early investments in child health. Chapter: The Effect of Caesarean Section for Babies in Breech Presentation on Child and Mother Health. Evidence from a Regression Discontinuity Design. PhD, Aarhus University.

YAN, J. 2010. Essays on Risky Health Behaviors and Policy Intervention. Ph.D., Washington University in St. Louis.

YAN, J. 2014. The effects of a minimum cigarette purchase age of 21 on prenatal smoking and infant health. Eastern Economic Journal, 40, 289-308.

YANG, M. Treatment Effect Analyses through Orthogonality Conditions Implied by a Fuzzy Regression Discontinuity Design, with Two Empirical Studies [Online]. Available: http://www.lehigh.edu/~muy208/research/rdd/rdd_wp_version.pdf [Accessed 18 November 2015].

YANG, M. 2008. Regression Discontinuity Design and Program Evaluation. Ph.D., University of California, Berkeley.

YANG, T.-T., HAN, H.-W. & LIEN, H.-M. 2014. Patient Cost-Sharing and Healthcare Utilization in Early Childhood: Evidence from a Regression Discontinuity Design. Canadian Centre for Health Economics.

YANOVITZKY, I., ZANUTTO, E. & HORNIK, R. 2005. Estimating causal effects of public health education campaigns using propensity score methodology. Evaluation and Program Planning, 28, 209-220.

YÖRÜK, B. K. & YÖRÜK, C. E. 2011. The impact of minimum legal drinking age laws on alcohol consumption, smoking, and marijuana use: Evidence from a regression discontinuity design using exact date of birth. Journal of Health Economics, 30, 740-752.

YÖRÜK, C. E. & YÖRÜK, B. K. 2014. Do Minimum Legal Tobacco Purchase Age Laws Work? : CESifo Group Munich, CESifo Working Paper Series: 4860.

YOU, J. 2013. The role of microcredit in older children's nutrition: Quasi-experimental evidence from rural China. Food Policy, 43, 167-179.

YU, B. & KAFFINE, D. T. 2011. Blue laws, DUIs and alcohol-related accidents: regression discontinuity evidence from Colorado. Journal of Economics (MVEA), 37, 21-38.

ZANI, C., TONINELLI, G., FILISETTI, B. & DONATO, F. 2013. Polychlorinated biphenyls and cancer: an epidemiological assessment. Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews, 31, 99-144.

ZHANG, J., HUANG, Y., WANG, X., LIN, K. & WU, K. 2015. Environmental polychlorinated biphenyl exposure and breast cancer risk: a meta-analysis of observational studies. PLoS ONE, 10, e0142513.

ZHANG, N. 2009. The determinants of children's health. Ph.D., Cornell University.

ZHAO, M., KONISHI, Y. & GLEWWE, P. 2013. Does information on health status lead to a healthier lifestyle? Evidence from China on the effect of hypertension diagnosis on food consumption. Journal of Health Economics, 32, 367-385.

ZHONG, H. 2014. The effect of sibling size on children's health: a regression discontinuity design approach based on China's one-child policy. China Economic Review, 31, 156-165.

ZHONG, H. 2015. Does a college education cause better health and health behaviours? Applied Economics, 47, 639-653.

ZIEGELHÖFER, Z. 2012. Down with diarrhea: using fuzzy regression discontinuity design to link communal water supply with health. Graduate Institute of International and Development Studies Working Paper.

ZUCKERMAN, I. H., LEE, E., WUTOH, A. K., XUE, Z. & STUART, B. 2006. Application of regression-discontinuity analysis in pharmaceutical health services research. Health Services Research, 41, 550-563.