Pharmacoepidemiology

using randomised control trials and observational studies in clinical decision-making

BJCP British Journal of Clinical Pharmacology

Pharmacoepidemiology: using randomised control trials and observational studies in clinical decision-making

Journal:	*British Journal of Clinical Pharmacology*
Manuscript ID	RU-00062-19
Manuscript Type:	Review - Un-commissioned
Date Submitted by the Author:	22-Jan-2019
Complete List of Authors:	Caparrotta, Thomas; Institute of Genetics and Molecular Medicine, Pharmacoepidemiology; University of Edinburgh Queen's Medical Research Institute, Clinical Pharmacology Dear, James; Edinburgh University, Centre for Cardiovascular Science Colhoun, Helen; Institute of Genetics and Molecular Medicine Webb, David; BHF Centre of Research Excellence (CoRE), Queen's Medical Research Institute, Clinical Pharmacology Unit
Key Words:	Health Policy < Clinical Pharmacology, Pharmacoepidemiology < Epidemiology, Quality use of medicines < Clinical Pharmacology, Evidence-based Medicine < Clinical Pharmacology
Abstract:	Weighing up sources of evidence is a key skill for clinical decision-makers. Randomised controlled trials (RCTs) and observational studies each have advantages and disadvantages, and in both cases perceived weaknesses can be improved through modifications of design and analysis. In the field of pharmacoepidemiology, RCTs provide better evidence than observational studies, largely because randomisation reduces bias and confounding. Although observational studies, even in a small cohort, can provide very useful clinical evidence, they may also be misleading (as shown by subsequent RCTs), in part because of allocation bias. There is an unmet need for clinicians to become well-versed in appraising the study design and statistical analysis of observational pharmacoepidemiology (OP) studies, rather like the medical training already offered for RCT evaluation. This is because OP studies are likely to become more common with the computerisation of healthcare records and increasingly contribute to the evidence base available for clinical decision-making. However, when the results of an RCT conflict with the results of an OP study, the findings of the RCT should be preferred, especially if its findings have been repeated elsewhere. Conversely, OP studies that align with the findings of RCTs can provide rich and useful information to complement that generated by RCTs.

Pharmacoepidemiology: using randomised control trials and observational studies in clinical decision-making

Dr Thomas M Caparrotta

Speciality Registrar, Clinical Pharmacology and Therapeutics

Diabetes UK 'Sir George Alberti' Clinical Research Training Fellow

University of Edinburgh, Scotland United Kingdom

tom.caparrotta@igmm.ed.ac.uk

Dr James M Dear

Reader, Consultant Clinical Pharmacologist and Toxicologist

University of Edinburgh, Scotland, United Kingdom

james.dear@ed.ac.uk

Prof Helen M Colhoun

AXA Professor of Medical Informatics and Life Course Epidemiology

University of Edinburgh, Scotland, United Kingdom

19 helen.colhoun@igmm.ed.ac.uk

20

21 Prof David J Webb

22 Christison Professor of Therapeutics and Clinical Pharmacology

23 University of Edinburgh, Scotland, United Kingdom

24 d.j.webb@ed.ac.uk

25

Key words

Pharmacoepidemiology

Clinical trial methodology

Health policy

Quality use of medicines

Evidence-based medicine

Word count

10,092

Abstract

Weighing up sources of evidence is a key skill for clinical decision-makers. Randomised controlled trials (RCTs) and observational studies each have advantages and disadvantages, and in both cases perceived weaknesses can be improved through modifications of design and analysis. In the field of pharmacoepidemiology, RCTs provide better evidence than observational studies, largely because randomisation reduces bias and confounding. Although observational studies, even in a small cohort, can provide very useful clinical evidence, they may also be misleading (as shown by subsequent RCTs), in part because of allocation bias. There is an unmet need for clinicians to become well-versed in appraising the study design and statistical analysis of observational pharmacoepidemiology (OP) studies, rather like the medical training already offered for RCT evaluation. This is because OP studies are likely to become more common with the computerisation of healthcare records and increasingly contribute to the evidence base available for clinical decision-making. However, when the results of an RCT *conflict* with the results of an OP study, the findings of the RCT should be preferred, especially if its findings have been repeated elsewhere. Conversely, OP studies that align with the findings of RCTs can provide rich and useful information to complement that generated by RCTs.

Introduction

Robust evidence about clinical interventions is necessary for many reasons, from the licensing of new treatments to informing clinical practice, guideline creation and clinical/cost effectiveness analysis. Pharmacoepidemiology involves the study of drug-based interventions in populations and for more than 70 years the randomised controlled trial (RCT) has been the mainstay of this field. RCTs differ from observational pharmacoepidemiology (OP) studies in one key way – the random assignment of participants to interventions. Randomisation serves to ensure that confounders and effect modifiers are randomly allocated between the groups, thus providing unbiased estimates of treatment effect. For this reason, they are the preferred approach for estimating relative and absolute treatment effects and therefore are more useful in supporting clinical decision-making. RCTs are most impactful from the epidemiological perspective where efforts have been made to increase their generalisability.

Observational studies have also provided valuable evidence in the field of medicine. They demonstrated the benefits of treating diabetes with insulin and the link between smoking and lung cancer, for example.[1,2] Indeed, observational studies remain universally accepted for delineating the natural history of diseases, their risk factors and prognostic markers. However, *observational pharmacoepidemiology*, where (beneficial/harmful) treatment effects are quantified, has been subject to criticism. This is because bias and confounding create difficulty in attributing cause and effect. It remains the case, however, that OP studies remain the mainstay of pharmacovigilance for harmful effects once a drug has been licensed.

A false 'conflict' between proponents of RCTs and OP studies has been created. Both types of study have important, often complementary, objectives and each can help deliver evidence not supplied by the other. Also, both RCTs and OP studies have their strengths and weaknesses and can provide flawed answers, through poor design, execution or analysis. Furthermore, there is increasing concern about the observed efficacy-effectiveness gap, which is the inconsistency between the effects of an intervention reported in clinical trials compared to that reported in

routine clinical practice. It is likely that well-designed OP studies (alongside more generalisable RCTs) will help to plug this gap.[3]

Well-designed and well-conducted RCTs have good internal validity, which in turn allows inferences on *(relative) efficacy* and causality to be made.[4] Efficacy is described as 'the performance of an intervention under ideal, controlled circumstances *compared to placebo*' whereas relative efficacy is similar to this except *comparison is to a standard alternative* rather than placebo. *Relative effectiveness* is described as an intervention's performance in a 'variety of endpoints important to patients and healthcare providers compared to the usual care offered by a health system in the population of patients identified as eligible for treatment by their care providers, subject to free and variable patient and clinician behaviour' and can be measured in pragmatic RCTs or in OP studies.[3,5]

Weighing-up clinical evidence is a key skill for clinical decision-making. The Academy of Medical Sciences has recently published an extensive report on the 'sources of evidence for assessing the safety, efficacy and effectiveness of medicines'.[6] In this article, we discuss the inherent properties, advantages and disadvantages of both type of study and how they might be improved. This will assist readers in balancing evidence to make clinical decisions, particularly in the field of OP, where robust methodology and statistical analysis is less well-understood. However, we argue that when the results of RCTs and OP studies in similar patient populations are conflicting, the results of a well-designed and executed RCT are more likely to represent an unbiased estimate of the treatment effect. However, well-designed and executed OP studies can confirm and extend the findings of RCTs and show that treatment works in groups often excluded from RCTs, such as older people, the very young and those with comorbidities.

Randomised controlled trials in pharmacoepidemiology

The inherent properties of RCTs make them the most robust means of evaluating healthcare interventions (see table 1 for brief description of study types).[4] When properly designed and executed, with sufficient power and appropriate analysis, RCTs give the best indication of the efficacy of an intervention.[4] The *key properties* of RCTs that differ from OP studies are:

- A pre-planned experiment, which gives rise to internal validity (and can reduce selection bias)
- Random allocation to treatment, which prevents allocation bias (also variously known as channelling bias, contraindication bias, confounding by indication, confounding by severity or confounding by frailty)
- Blinding, which avoids observer bias (although some RCTs are not blinded)

The advantages of RCTs stem from:

- The reduction of bias and the equal distribution of confounders and effect modifiers provided by randomisation
- Blinding, when done, reduces observer bias and contamination of result reporting by patients knowing which intervention they have received, particularly when outcomes are subjective
- Formal calculation of adequate trial size to ensure satisfactory study power and thus meaningful results
- The minimisation of missing data and the systematic collection of outcomes to prevent information bias (see table 2 for a description of biases and their mitigation)

An effect modifier is a clinical characteristic (e.g. age, sex, genotype etc.) which causes the *effect* of the exposure to change (e.g. hormone replacement therapy's protection from endometrial cancer only appears to operate in women with a body mass index (BMI) >30, thus in this context BMI can be considered an effect modifier).[7] Bradford Hill lists a number of criteria that increase

confidence that an association is causal (see table 3 for these criteria as applied to the medical sciences). He states that 'experimentation' lends the strongest support to causality – the design of RCTs can fulfil the experimentation criterion and support causal inferences.[8]

RCTs also have limitations, assuming otherwise robust design. These relate particularly to the generalisability of results (i.e. who gets selected for inclusion in the study and whether they are representative of the population to whom the results will be applied). Other limitations of RCTs relate to the length of follow-up, which when long can dramatically increase costs, and trial size, which can also increases costs and when inadequate can mean insufficient power of the trial to detect treatment effect and, more commonly, rare safety event outcomes.

If an RCT is improperly designed, performed or analysed it may mislead more than a well-designed OP study that attempts to account for bias and confounding.[9] In the following sections the characteristics of RCTs in pharmacoepidemiology and strategies to ensure their good conduct are addressed in more detail.

The advantage of randomisation, allocation concealment and blinding

Randomisation is a major contributor to the benefit RCTs have over observational studies. The random allocation of participants to treatment groups achieves comparability between these, especially in terms of prior risk of the outcome of interest and any effect modifiers (see table 2 for a brief overview of biases and their mitigation). This allows causal inferences to be made; in other words, the treatment effect observed is likely due to the intervention, all things being equal. If certain co-variates might not be equally distributed between treatment groups with patient-level randomisation then *stratified randomisation* might be employed to improve group comparability, e.g. it might be important there be equal numbers of patients with a rare, severe disease phenotype in both arms.[10] In either case, whether using stratified or patient-level randomisation, any increase in comparability applies equally to variables we can and cannot

measure, reducing or balancing confounders and effect modifiers.[10–12] It is essential that the randomisation process is not compromised, which is achieved through robust randomisation methods and allocation concealment.[9]

Importantly, allocation concealment and blinding of allocation are not the same. The former means hiding the sequence of allocation prior to recruitment, so it is not possible to predict to which treatment group a participant will be assigned.[13] Blinding is the process of continuing allocation concealment until the end of the study and is easier to do in RCTs than other types of epidemiological study.[14] The effect of blinding is to reduce observer bias in ascertaining the outcomes of interest, a form of differential information bias (see table 2, for a more in depth description of information bias).[15] . Ideally, all participants and staff ought to be blinded, but this is not always practicable.[15,16] In RCTs, although blinding requires allocation concealment, allocation concealment is not always followed by blinding. When allocation concealment is undertaken but the study is not blinded, a trial is said to be open-label.

Open-label studies are thought to increase the risk of observer bias. To minimise this, in open-label studies, staff analysing the *outcome* data should be blinded to allocation, as this is almost always possible, and is particularly important when the outcome is subjective.[13]

PROBE: A particular type of open-label study design, the Prospective Randomised Open Blinded End-point study has been proposed, thought to be more cost-effective than the double blind prospective study. The study design uses strict randomisation and hard end-point definitions (ones that are well-defined and measured objectively) to allow for the comparison of interventions to take place. Outcome data are collected through routine clinical care which may increase the generalisability of PROBE study findings. However, the open-label nature of the trial may introduce observer bias (a type of information bias) in the recording of the end-point, even though the use of hard end-points tries to reduce this possibility. Also patients know which intervention they are exposed to, which introduces the risk of contamination if they seek a different treatment from another healthcare provider that goes unrecorded in the trial.[17]

Reduction of bias and confounding in RCTs compared to OP studies

Bias is 'a systematic (as opposed to random) distortion, due to a design flaw, interfering factor or judgement that can affect the conception, design or conduct of a study or the collection, analysis, interpretation, presentation or discussion of outcome data, causing erroneous over-/under-estimation of the probable size or direction of a treatment effect or association'.[18,19]

Confounders are extraneous factors directly, or inversely, associated with the variables being measured such that it creates the false impression of an association between the variables which, were it not for the confounder, does not exist.[20]. Confounders differ from biases in that, if the confounder is known, statistical methods can be employed to adjust for its effect at the analysis stage, which is not always the case with bias as it cannot be corrected for once introduced into a study.[21] It is of course, not possible to correct for unknown confounders. Spurious associations are associations that are not real and occur due to chance and bias. Indirect associations are real but not causal and stem from confounding.

The design of RCTs can significantly reduce both bias and confounding and hence spurious and indirect associations. In the main, there are three categories of bias which the design of an RCT minimises (indeed most biases fit into one of these broad categories, despite their varying nomenclature); selection bias, allocation bias and information bias. In *selection bias* there are systematic differences between those who are *observed (included) in a study compared to those who are not*, particularly in terms of prior risk of the outcome of interest or effect modifiers. This means that the population under study is no longer representative of the condition being investigated and participants differ from the population to whom the results are to be applied *independently of the interventions being studied*. *Allocation bias* occurs when there are systematic differences between how participants *are assigned to their treatment group*, meaning that those exposed to an intervention differ from those not exposed in terms of their prior risk

of the outcome of interest or effect modifiers. *Information bias* occurs when information is obtained differently (for whatever reason) between the groups such as a flaw in measuring exposure, outcomes or co-variates with differing accuracy, intentionally or unintentionally. See table 2 for more information about biases and their mitigation. Randomisation, with a large enough sample, operates to reduce confounding and differential effect modification, by balancing these between the groups.[20]

The Cochrane Handbook for Systematic Review of Interventions contains a tool which uses readers' judgement for assessing the risk of bias in a study, hence making a verdict about its internal validity and in turn whether the study merits inclusion in a systematic review.[22] In order to maintain the benefits inherent to RCTs and provide for adequate reporting of protocols and results, the following paragraphs describe agreed reporting standards.

CONSORT: The CONsolidated Standards Of Reporting Trials [23] aims 'to alleviate the problems arising from the inadequate reporting of RCTs'. It consists of an evidence-based minimum standard of recommendations to assist with complete and transparent reporting of RCTs, thereby aiding critical appraisal and interpretation. Since its inception in 1993, CONSORT has undergone regular revisions; the current 2010 revision consists of a 25-item checklist and flowchart focussing on how the trial was designed, analysed and interpreted; a central tenet is the pre-registration of trial protocols. In particular, adherence to CONSORT may reduce selective reporting bias (a type of information bias) and allows the reader to ascertain whether included analyses were pre-planned or not and, if not, to offer an explanation as to why the analysis became necessary. A number of studies have investigated the effect of the 2001 revision to the CONSORT guidelines on the completeness of reporting.[24–26] Although these found that there was a general trend of improvement in the reporting of important aspects of trial methodology, it remained sub-optimal.[24–26] Endorsement of CONSORT by journals may beneficially influence the completeness of trial reporting in the medical literature.[24]

Statistical analysis plan (SAP): this is a critical document to the undertaking of RCTs (and indeed OP studies) and making the plan available supports transparency and reproducibility, especially since statistical decisions carry great weight for the conclusions of a trial. SAP is a 'more detailed and technical elaboration of the principle features of analysis included in the trial protocol'[27]. However, until recently (2017) no guidance existed for SAP contents (compare with CONSORT which has existed since 1993). A recently published expert consensus document has now specified minimum content for SAP in relation to RCTs[28] (and also now exists for OP studies[29]). It will be important to measure whether this guideline improves the transparency of reporting of statistical analysis and consequently whether this improves the reproducibility of RCTs (and OP studies).

RCTs have internal validity, which allows causality to be established

Well-designed and well-conducted RCTs have internal validity, which is the extent to which causal conclusions regarding a study are justified.[30] This is especially the case when the population being examined is large and, by analogy, if the findings are replicated in other studies. A trial with good internal validity is likely to have true results for the population with the characteristics being studied; in other words, any effect detected is likely to be *caused* by the treatment.[12] However an RCT may produce a treatment effect estimate that may not be generalisable beyond the population being studied, despite having good internal validity. Conversely, an RCT with good generalisability has a treatment effect estimate which can be applied more broadly. However, in order to calculate an absolute risk reduction (ARR), the treatment effect estimate obtained must apply to the background population against whom the ARR is to be calculated.

Bradford Hill lists a number of criteria which increase our confidence that an association is causal (see table 3 for these criteria as applied to the medical sciences).[8]

RCTs facilitate the comparison of treatments

Due to their results likely representing the truth, RCTs can *directly* compare different treatments head-to-head when an active comparator control is used instead of placebo. This allows conclusions regarding *relative efficacy* to be made. Multiple-arm studies can be used to demonstrate non-inferiority or superiority, comparing multiple treatments or dosages simultaneously, and are becoming more common.[31]

However, with an ever-increasing number of treatments, it will be impractical to carry out head-to-head comparisons of all treatments available. A well-conducted RCT – due to the confidence that the treatment effect observed is likely to be true, because of its internal validity - can more easily be incorporated into adjusted indirect comparisons, mixed treatment comparisons, meta-analyses and systematic reviews than OP studies. RCTs allow statistical inference to be made regarding the efficacy of different interventions even when head-to-head comparison has not been undertaken.[32]

Maximising the results of RCTs in pharmacoepidemiology

In this section, aspects of trial design and conduct that are aimed at maximising the validity and reducing the impact of the constraints inherent to RCTs are explored.

Although RCTs are preferable for assessments of efficacy, individual trial methodology must be scrutinised in order to critically appraise the results, see table 4 for factors that might reduce confidence in RCTs or meta-analyses combining RCT results.

The findings of RCTs may not be generalisable

RCTs are often conducted in select groups of patients, in specialist centres, by leading experts, using state-of-the-art technology for a limited period of time, so-called explanatory RCTs – unrepresentative of the care that patients receiving the intervention in the community would experience. While these trials are critical in establishing efficacy and preliminary safety it may mean that the study results are only valid in the specific group of participants included in the trial. Also, some RCTs have been criticised for using surrogate or composite endpoints which do not translate into clinical benefit, or for not taking into consideration factors important to patients' well-being.[33]

Generalisability, relates to the degree to which a treatment effect estimate can be applied to a wide group of patients under 'usual conditions' (effectiveness/external validity - see table 5 for some factors to consider when assessing generalisability). This may not apply to an RCT unless steps are taken to make the study more generalisable, such as using some of the following methods.[33]

Intention to treat analysis (ITT): means that participants are analysed in the group to which they were assigned, irrespective of whether they completed the study. This can help prevent attrition bias.[34]. Attrition bias (loss to follow-up) can threaten the internal validity of RCTs by removing the benefits of randomisation, introducing the potential for bias, confounding and an imbalance of the prior risk of the outcome of interest between the study groups (and also reduce the study's power). Thus, attrition bias is a form of *after the event* selection bias where one group (or both) are no longer being representative of the condition under study.[33,34]

ITT analysis prevents biased estimates of treatment effect due to differential drop-out between treatment arms and sensitivity analyses should be presented comparing relative treatment effect estimates of per protocol results (only those who fully completed the study protocol) to ITT results. ITT evaluation might also better reflect real-world clinical practice (*increasing generalisability*), where patients may stop the intervention or adhere poorly to it and thus gives a more realistic estimate of the treatment effect in the wider population (as these non-adherent patients are accounted for) although the maximum achievable benefit may well be underestimated.[10,20] It is also good practice to provide the sub-group characteristics of the patients who were lost to follow-up.[34]

Pragmatic RCTs: (pRCTs) aim to redress perceived problems in the generalisability of RCTs by providing answers to clinical problems relevant to patients and clinicians.[35] Often pRCTs aim to investigate heterogeneous patient groups, may not employ placebos and use outcome measures which might include return to work, reduction in general practitioner visits and quality of life, in addition to outcomes related to efficacy.[35] An example of a pRCT is the Salford Lung Study, which randomised ~50% of the patient population in the community, and demonstrated the superiority of fluticasone furoate and vilanterol over usual care in the management of COPD, the results of which are broadly generalisable.[36]

Large simple RCTs: (lsRCTs) are pRCTs but with protocols mandating only minimal data collection on outcomes important to patients or care providers. They are well-suited to

assessing outcomes which are rare or have long latency, when study populations are heterogeneous or when many risks need quantifying. lsRCTs minimise the complexity and volume of data analysis as outcomes are measured from routine care. An example includes a trial demonstrating that ziprasidone is not associated with an excess of non-suicide mortality, despite being associated with QT-prolongation on the electrocardiogram.[37]

Randomised database studies: are a specific form of lsRCT, capitalising on the data held in electronic healthcare records or disease registry databases. They attempt to achieve both internal and external validity although the optimal approach to important issues like participant consent are still to be standardised. This type of trial is likely to become more common with the increasing availability of computerised health data. An example includes a Swedish study which demonstrated that thrombus aspiration prior to stenting in acute ST-elevation myocardial infarction was no better than stenting alone, with similar outcomes in all sub-groups.[38]

Other approaches for improving generalisability: when combined into meta-analyses, the results of many RCTs can be made more generalisable by merging and comparing the different study populations investigated in the individual trials.

RCTs can be expensive and difficult to undertake

RCTs can be expensive to conduct. This is in part due to the exacting standards required by the study protocol to ensure internal validity and safety and also the length of time required for follow-up.[10] Also, the burden of regulation provided by internationally-agreed documents such as the International Conference on Harmonisation of Good Medical Practice (ICH-GCP) are viewed by some as an impediment to speedy research. This is because the burden of specific data collection and retention standards for trials concerned with the supply of data for the registration of interventions for human use. [39,40] Costs can increase further if larger groups of patients are needed to power RCTs.[10] In addition, in some situations, the outcome of interest

may be so far in the future that it is difficult and expensive to maintain follow-up.[41] Furthermore, the elevated levels of monitoring, compared to routine clinical care, in order to meet regulatory requirements and data capture can add significantly to the cost burden of running RCTs.[42]

Cluster RCTs: (cRCTs) randomise at the group-level, say a clinic or hospital, rather than at the individual patient-level. This type of study design can significantly reduce costs by reducing the administrative burden of the trial since changes are introduced wholesale at the group-level, do not require individual patient-level consent and may also be more easily deployed in emergency situations. Deciding the unit of inference (whom the trial results will apply to) early is essential in the study design to prevent the occurrence of ecological fallacy (drawing individual conclusions from group-level data or *vice versa*). This type of trial has less statistical efficiency than randomising an equivalent number of people at the individual-level. [43] An example of a cRCT is the Randomised Evaluation of an Algorithm for Crohn's Treatment Trial, which randomised gastroenterology clinics to either standard incremental therapies for disease control or to early combined immunosuppression and demonstrated that there was no different in primary outcome between the units of randomisation.[44]

It may be impossible to undertake an RCT in emergency situations; for example following a terrorist incident or during an epidemic, where randomisation may be impractical or unethical, and where there is a need to produce information quickly while at the same time minimising the risk to patients and staff.[45] The cRCT can help with these perceived difficulties. The 2014-2015 Ebola outbreak led to the design of novel approaches to undertaking RCTs. The adaptive ring vaccination, open-label, cluster RCT (Ebola 'Ça Suffit' Trial) used to demonstrate efficacy of rVSV-vectored Ebola vaccine, where immediate versus delayed vaccination was compared and immediate vaccination was favoured.[46] In ring vaccination, at-risk patients are identified for vaccination by being contacts of a known Ebola case and had been used successfully during smallpox eradication, but not before as a clinical trial methodology.[47]

Crossover RCTs: (xRCTs) can give greater precision of treatment effect, given the same number of subjects, than a similarly sized parallel group study. xRCTs are a within-subject study design, where participants are randomly exposed to an interventions in sequence (treatment A followed by treatment B or *vice versa*), and thus act as their own controls.[48] One of the treatments may be placebo or an active control. This reduces inter-subject variability (thus increasing precision) but are not appropriate if there is a significant carryover effect from one of the treatments, despite a washout period between treatments.[48,49] xRCTs can thus increase study power but cannot be employed for conditions with an acute natural history nor investigate treatments that provide a cure rather than respite.[48] For example, an xRCT was used to investigate sequential plasmapheresis versus sham plasmapheresis (placebo) treatment in the same patients on symptom improvement in rheumatoid arthritis and failed to demonstrate a difference between the treatments.[50]

Factorial RCT: (fRCT) is another design that can assess outcomes more efficiently than separate trials.[51] This allows for the assessment of multiple treatments in the same population, maximises study power and also provides information on interactions between treatments.[51] In its simplest form, a '2x2' fRCT, say treatments A or B and C or D exist. This fRCT would allow the comparison between treatment A and C or D, or treatment B and C or D. An fRCT can help explain which treatment is better, either alone or in combination and whether or not there is a synergistic or additive effect between treatments.[51,52] An fRCT was used by investigators to assess whether a shortened course of *N*-acetylcysteine (NAC) in paracetamol intoxication was associated with fewer side effects than a 21-hour course, while at the same time assessing whether pre-treatment with the anti-emetic ondansetron reduced nausea and vomiting due to NAC treatment. The study found that both the shortened course of NAC and ondansetron pre-treatment reduced nausea and vomiting independently and also additively (and that the shortened course of NAC was associated with fewer anaphylactoid reactions).[53]

Issues related to study power and lower than expected event rate in RCTs

Careful thought is given to the planning of RCTs, in particular to ensure internal validity and adequate power. However, should an insufficient number of participants be recruited, more participants than expected drop-out or a lower than expected event rate be observed, then the trial may not have sufficient power to detect significant change or may have to continue for longer than planned to yield an adequate effect estimate.[54] Some of the study designs detailed below can help address these issues, in addition to the advantages already described by using large pRCTs and lsRCTs.

Multi-arm studies and adaptive clinical trials: may be superior to two-arm studies, which frequently fail to demonstrate superiority.[55] Patients, clinicians and regulatory authorities want to know whether certain interventions are superior to those already offered as quickly as possible.[55,56] Multi-arm superiority trials allow the direct comparison of many different treatments or different treatment regimens compared to an active comparator group. They are simpler, quicker and cheaper than a series of two-arm trials investigating the same and provide data for direct comparison rather than many two-arm studies being compared in meta-analysis, which causes difficulties in interpretation when the studies are heterogeneous.[55,56] It may also be the case that multi-arm trials recruit more effectively than two-arm trials, possibly since the multiple arms, with different inclusion criteria, means more patients are eligible, and well-designed multi-arm studies may provide significant patient benefit compared to multiple two-armed trials.[55] The ideal study design is yet to be established and the multiple arms of the study may cause difficulties in the interpretation of results, particularly when arms are removed or added.

Adaptive clinical trials are a type of trial design, with multiple arms, which seek to address some of the concerns about multi-arms studies. To-date, they have mainly been deployed in the field of oncology, since their design can handle the increasingly-recognised biological heterogeneity of tumours. However, they show promise in other fields such as neurological degeneration. There are different forms of these trials, but all allow some form of pre-specified adaptation, guided by a master protocol, to take into account evolving understanding both from within and outside the adaptive trial. They aim to address the perceived deficiencies of traditional trial designs, which has been described as the 'weakest link' in cancer therapy development now that molecular understanding of tumour biology has increased.[57] The STAMPEDE trial investigating treatment alternatives for advanced prostate cancer is one example of an adaptive, multi-arm, multi-stage platform trial.[58] However, the multiple comparisons between arms, particularly if new treatment arms are added or removed, and the necessity to undertake repeated interim analyses make these trials extremely complicated to analyse and interpret, as the multiple comparisons increase the likelihood that a positive finding is significantly different by chance alone if the multiple comparisons are not accounted for.[56,59]

In *basket trials*, eligibility is determined by a master protocol often defined by the presence of a molecular alternation rather than a specific tumour site. Each basket represents a molecularly-defined sub-trial (drug-mutation pair testing) with matched therapy or control. In *umbrella trials* a single class or type of tumour is molecularly screened and assigned to sub-trials in light of these results, where the molecular signature refines rather than defines inclusion (compare with basket trials, where inclusion is defined by the molecular signature). *Platform trials*, have a common control arm but many different experimental arms that enter or exit the trial as effectiveness or futility are demonstrated (often according to Bayesian decision-making rules). Adaptive randomisation, where patients with a particular molecular signature are preferentially enrolled into the trial arms that show the most promise, may also be a feature.[57]

Stopping after pre-specified number of events: some trial designs, particularly when assessing event-based outcomes, power the trial in terms of a minimum number of primary outcome events to be observed, rather than pre-specifying the number of participants to be recruited or the length of time under observation (although event rate is influenced by these two latter parameters). This may increase trial efficiency by being able to declare a treatment effect as significant on the basis of an observed difference in event rates between trial arms, without having to wait for a pre-specified number of participants to be enrolled or for a specific length of time to have elapsed. This type of design allows for flexibility should the event rate assumptions in the trial design be greater than the number of events observed in practice, which would otherwise result in the trial being under-powered. The cardiovascular outcome trial examining canagliflozin for the treatment of type 2 diabetes (CANVAS programme) adopted this approach to demonstrate that canagliflozin reduced the number of major cardiovascular outcome events compared to placebo.[60]

The advantages of pre-specifying sub-group analysis

RCTs often report results with regards to sub-groups differing from each other in baseline traits. Trial populations are often heterogenous, which raises the question of whether effects observed are true for all of the patients regardless of their baseline characteristics.[61,62] If conducted appropriately, sub-group analysis can be illuminating, increasing generalisability, and impacting positively on patient care through improved clinical decision-making.[61] However, if sub-group analysis is performed poorly, or indeed not reported, it can be misleading.[61,62] Pre-planned sub-group analysis forms a key part in all the published criteria designed to help readers decide whether the sub-group effect is real and is also encouraged in CONSORT.[23,61] However, systematic reviews have shown that reported sub-group analyses are seldom pre-specified and there is a recognition that uncontrolled flexibility in the analysis of data carries a real risk of

false positive findings.[61–64] If multiple assumptions are tested in sub-group analysis the likelihood of a falsely significant result by chance alone increases.[65] Although unscheduled sub-group analysis, labelled as such, may have a role in hypothesis-generation for subsequent trials, statistically inferential approaches to sub-group analysis should be limited to small numbers of pre-specified sub-groups underpinned by sound biological evidence to limit the reporting of false positive effects.[66]

The conflicting tensions in stopping trials early

A complex problem is the early termination of RCTs due to *beneficial effects* becoming apparent, where there is a tension between obtaining a true estimate of the treatment effect and denying potential users a beneficial new treatment. When a trial is stopped early its internal validity is compromised. It has been suggested that a trial stopped early for beneficial reasons may *over-estimate* the treatment effect because of the following: the decision to stop trials early requires that data be analysed on multiple occasions; probability theory means that the more times the data are analysed, the more likely the data will yield a 'random high' leading to the trial being stopped (which would not occur if the reading were not 'high').[30,65,67,68] Stopping trials early also reduces the likelihood of adverse effects being detected as there is less time for these to accumulate. Methods such as 'increasing nominal significance' for each analysis (e.g. the O'Brien-Flemming method), which raise the threshold for stopping at interim analysis, can lessen the risk of 'random highs' leading to trial termination by raising the threshold for this to happen and should always be *pre-specified* in the SAP.[69]

Trials may also be stopped early for *futility*, which is the inability of a clinical trial to meet its objectives.[70] On the one hand stopping trials early for futility protects participants from being exposed to ineffective treatment and can save resources to be spent of more encouraging research. On the other hand, stopping for futility may leave secondary research questions

unanswered and trials that fail conventional significance testing may still be consistent with a probable positive effect and contribute to the total evidence (in meta-analysis).[71] Futility rules must be considered *before starting a trial* and always be included in the SAP although it may not be clear *a priori* how to choose the most appropriate stopping boundary. Indeed, many trials are continued to their conclusion unless there is clear evidence of harm. Various statistical methods exist for assessing futility including conditional rules that attempt to calculate the ultimate likelihood of success. Some of these may be unduly influenced by early participants in the trial.[70–72] The problem of stopping trials early for futility risks the opposite effect to stopping trials early for benefit, where a random high causes the trial to be terminated. In stopping for futility, early results may represent 'random lows' which cause the illusion of no effect leading to the trial being stopped, when in fact, had more information been gathered this 'no effect signal' would disappear.

Assessment of low frequency or long-term harms

RCTs are not often powered to detect harms that occur infrequently or that only develop a considerable time after exposure.[73] Indeed, RCTs do not often identify a pre-defined hypothesis for harms, which contrasts with the usually reported pre-specified hypothesis for efficacy.[73] In addition, in explanatory RCTs, due to their strict inclusion and exclusion criteria, patients at the highest risk of being harmed are often excluded (e.g. older patients, those with multiple comorbidities, children), although they may become the main users of the treatment if licensed.[73] Furthermore, in longer-term, larger RCTs it can be challenging to distinguish harms which are caused by the treatment (iatrogenic) from those which are 'inter-current and non-causal or just random error'.[65] It should be noted that it is recognised that rare harms may not become apparent until after a therapy has been licensed and table 6 illustrates the number of patients to be observed to detect a given adverse event rate. In meta-analysis or systematic

reviews, conclusions about harms may also be misleading if the available data are affected by publication bias.[73]

While it is important to consider the ITT effect estimate when examining *efficacy* of an intervention, it is important to note that, for *safety evaluation*, on-treatment (and per protocol) analyses should be considered. This is because any harms caused by an intervention are more likely to occur in those exposed to the treatment than those who are not.

Observational Pharmacoepidemiology

Observational studies (see table 1 for brief description of study types) include cohort, case-control and cross-sectional studies.[74] Cohort studies can be prospective or retrospective, with individuals exposed to an intervention identified, compared to non-exposed individuals, and any difference noted in the event rate over time.[74] Case-control studies are retrospective. Cases are identified after an event has occurred, compared to similar controls in whom the event has not occurred and any differences in exposure established retrospectively.[74] Cross-sectional studies look at the prevalence of a disease at a specific time point and may use historical data to establish exposure.[74]

The key difference between these types of study and an RCT is that, in OP studies, the intervention is selected for/by a patient, or the patient is selected on the basis of having been exposed to the intervention, rather than the intervention being allocated randomly.[74] This makes it conceptually more difficult to attribute an outcome to a particular treatment and also introduces the potential that biases or confounding account for any differences observed.[74] In particular, an extremely challenging problem in OP is *allocation bias*. In this bias, patients assigned a therapy are different from patients not assigned the therapy (e.g. more severely unwell participants treated more often with the intervention compared to less severely unwell patients) meaning the any difference in treatment effect observed may be due to this difference between patients and not the treatment. Also, the sensitivity and specificity of the outcome measures are often unknown, so it is unclear whether all outcome data have been captured and in what depth of detail, which leads to information bias.

Despite these perceived deficiencies, OP studies (as cohort or case-control studies, but more often as an adverse event reporting system e.g. the Yellow Card Scheme in the UK) are an important part of post-licensing pharmacovigilance. With the increasing availability of electronic health records claims databases and disease registries there is renewed interest in OP studies for making inferences on the effectiveness of interventions as well as quantifying

potential harms. Although most clinician are well-trained in assessing the validity of RCTs, there is less widespread knowledge of appropriate study design and statistical methods for OP. However, it is vital for healthcare professionals to become versed in OP study appraisal as an increasing number of these studies are likely to be published in the future, given the increasing accessibility of computerised observational data and a strong push to harness these data.[3,6,75] Clinicians will need to understand whether the study design used is appropriate given the question and whether the data analysis methods are robust enough to have confidence in the results.

Clinical Pharmacologists are particularly well-placed to be at the forefront of robust OP study production given their training in drug discovery, mechanisms of drug action, stratified pharmacology and drug safety. Indeed Clinical Pharmacologists are already producing informative research by conducting studies underpinned by sound biological principles such as the cohort study demonstrating that paroxetine use was associated with an increased risk of death in women with breast cancer treated with tamoxifen (paroxetine inhibits cytochrome P450 enzyme 2D6, which in turn reduces the bioactivation of tamoxifen necessary for its clinical effect).[76]

The following sections address in more detail the characteristics and strengths of OP studies in and strategies to ensure their good conduct.

OP studies allow quantification of effectiveness and can have good external validity

OP studies are often said to have high external validity, in that they are conducted in the 'real-world' and often have high levels of heterogeneity in their study populations making their findings more generalisable.[3,12,77] An OP study might confirm an intervention as effective in a heterogeneous population sample, when the intervention has previously been demonstrated as efficacious in an RCT. This is especially the case when the OP study includes some similar

participants to the RCT demonstrating efficacy and if the treatment effect detected in this sub-group of the OP study is in the same direction and order of magnitude as that reported in an RCT.[77] As such, OP studies can confirm and broaden the findings of RCTs to a wider population.[78]

STROBE: STrengthening the Reporting of OBservational studies in Epidemiology, aims to 'reduce the incomplete and inadequate reporting' of data in observational studies, 'which hamper the assessment of strengths and weaknesses of the studies reported in the medical literature' and to 'improve the quality of reporting'.[79] Rather like CONSORT, STROBE consists of guidelines and a checklist of 22 items considered essential for good reporting.[79] The current guidelines date from 2009, which is the first iteration and remains current. They cover the three most commonly employed designs in observational studies: i) cohort studies, ii) case-control studies and iii) cross-sectional studies.[79] Also like CONSORT, adherence to STROBE may reduce bias and allow the reader to ascertain whether included analyses were pre-planned or not, and if not to offer an explanation as to why the analysis became necessary.[79] STROBE came into existence more recently than CONSORT (2007 vs. 1996) and thus there is less evidence to suggest that it improves the quality of reporting although a bibliographic study found that of the observational studies analysed, over 80% made appropriate use of STROBE.[80]

OP studies can be carried out over a long period of time, detect rare adverse events and have lower costs

Observational studies can be carried out over much longer periods of time than RCTs. Indeed, there are cohort studies that have been running for many decades, such as the Framingham Cardiovascular Cohort Study, which has been running for over 65 years.[81] This advantage of time means that observational studies are able to provide important data on patients' long-term

longitudinal experiences, particularly in the setting of chronic diseases with a natural history over many years.[82]

RCTs are often not sufficiently powered to detect adverse events that occur very infrequently. For instance, to detect a doubling of an event rate from 0.1% to 0.2%, ~50,000 participants would need to be studied in an RCT to achieve an 80% power of detecting this at a p value of 0.05.[83] The extended period over which OP studies can be undertaken and the relative ease of obtaining a large enough population sample, compared to RCTs, makes OP studies suited to the defining of adverse events and their incidence.[82] Indeed, OP studies are an integral part of the post-marketing surveillance programme of newly approved drugs (e.g. adverse event reporting systems) and are occasionally mandated by regulators if there is an inconclusive safety signal in pre-licensing RCTs.[83] Observational studies can facilitate the detection of rare (<1/1,000) and very rare (<1/10,000) adverse events and are also able to provide long-term data on an intervention's tolerability.[78,84] See table 6 summarising the number of patients to be observed to detect a given adverse event.

Since observational studies frequently run in parallel with routine clinical care, they are often associated with lower costs than RCTs.[85] In addition, OP studies might employ data available from clinical databases such as the Clinical Practice Research Datalink (CPRD) in England, the Scottish Care Information - Diabetes Collaboration (SCI-DC) database and Health Maintenance Organization Research Network (HMORN) in the United States.[86–88] An ancillary benefit of these lower costs is that OP studies may include much larger numbers of patients.[85] Indeed, the future of OP is likely to be represented in such longitudinal electronic healthcare records (disease registry or insurance provider) databases.

OP studies can provide data to justify RCTs

636 OP studies often provide the evidence to justify an RCT.[89] Observational studies are also often
637 used in 'hypothesis generation' for RCTs (see table 7 for areas which are suited to observational
638 studies). In addition, if the treatment effect detected in an OP study is very large then it is not
639 always necessary to undertake an RCT. [89] There are multiple examples of treatments becoming
640 established on the basis of observational data without confirmation in an RCT, such as, for
641 example, the treatment of type 1 diabetes mellitus with insulin.[2]

642

643

Maximising the results of OP studies

The perceived disadvantages of observational studies in pharmacoepidemiology are discussed below alongside methods available to diminish these, related to both study design and methodology.

Bias and confounding make causality more difficult to establish in OP studies

The non-random allocation of patients in OP studies means they are more prone to biases and confounding, both known and unknown.[90,91] Although there may be strategies to mitigate the effects of biases, such as matching and multi-variable regression analysis, it is never possible to correct the results for all of these possible influences, in particular the ones not known about (see table 2 for a list of biases and their mitigation). Bradford Hill lists a number of criteria which increase our confidence that an association is causal (see table 3 for these criteria as applied to the medical sciences).

Due to the inherent difficulty controlling for bias in OP studies, causality is more difficult to establish since statistical association does not imply causality. It has been suggested that the stronger the treatment effect found in OP studies the greater the support for causality and this support is stronger still if the observation of association is consistent in different populations and with different study designs. [65,90–92]

OP studies can lead to inflation of positive treatment effects and under-estimation/under-reporting

667 The failure to use random allocation and blinding can be associated with relative increases in
668 the estimate of effects. Indeed, the distortion caused by not randomising and blinding during a
669 study has been associated with effect estimates as large or larger than the treatment effect
670 itself.[16] On the other hand, meta-analyses of the treatment effects in OP studies and RCTs have
671 demonstrated that, by and large, when good quality studies are analysed, the direction and
672 magnitude is broadly similar.[93–95] Nevertheless, the spectre of treatment effect over-inflation
673 hangs over OP studies and should always be borne in mind when considering their results.

674 Many OP studies rely on data gathered through routine clinical practice. Conversely, for
675 different reasons to those just described, this means that OP studies may also be at risk of
676 under-estimation, where a patient fails to seek healthcare and thus the true incident rate of a
677 condition may not be recorded, or under-reporting, where following interaction with a
678 healthcare system the data are inadequately reported. This under-estimation is a form of non-
679 differential information bias.[96]

680

681 ***Approaches to dealing with the limitations of OP***

682

683 Different methods (in terms of study design, analysis or both) have been employed to reduce
684 the effect of bias and confounding in OP, some of which are only appropriate in very specific
685 circumstances. One rule of thumb as a validation method is whether, within the OP study, a
686 group of subjects meeting the inclusion and exclusion criteria of a published RCT exploring the
687 effect of the same drug can be discerned. If it can be demonstrated that the patients in this sub-
688 group have a treatment effect detected that is in the same direction and of the same order of
689 magnitude as that found in the RCT, then this increases confidence that the treatment effect in
690 the larger, more heterogeneous group of patients is robust.

691

Study design and analysis methods to reduce bias and confounding in OP

Incident-user design: is a cohort study design aimed at reducing allocation bias, where incident-users (new users) of a drug for a particular indication are compared to incident-users of a different drug (controls) for the same indication. This assumes that *both users and controls* have been identified by clinical staff as benefitting from a new prescription and makes users and controls more similar, particularly in characteristics which may not be observable.[97] However, this does not always mean that incident-users and their controls are identical – for instance clinicians may avoid prescribing newly licensed drugs to frail elderly patients, sticking instead to drugs they are more familiar with using in this group. In this case the users and controls would cease to be as similar. Incident-user design also means precluding prevalent-users (longer-term users) from the study, reducing sample size and losing potentially valuable information. This design can be modified for the investigation of second- or third-line treatments by examining those that switch/add treatment for the same indication, as this switching/adding is not a random event, but rather influenced by disease worsening or a side effect again believed to improve comparability between 'switchers'.[97]

Natural experiments: are alternatives to RCTs which utilise naturally occurring circumstances to separate variables that usually associate together in a before and after cohort study.[63] One example is *universal exposure* to avoid selection and allocation bias, where the exposure occurs in total populations rather than through choice, and thus allows comparisons to be made between exposed and unexposed populations and causal inferences to be made.[98] This was the case in Japan, when use of the measles, mumps and rubella (MMR) vaccine abruptly stopped due to concerns about cases of aseptic meningitis, perhaps caused by the particular strain of mumps virus used to make the Japanese vaccine. This allowed the exploration of whether the MMR vaccine was associated with regressive autism, when concerns about this association surfaced a number of years later. Here, analysis of the Japanese population before and after the

cessation of widespread MMR use found no link between MMR and regressive autism.[99] Another example of natural experiment, devised as an alternative to RCTs, albeit applicable in limited circumstances, is *regression discontinuity design*. This uses a pre-determined 'assignment variable' (e.g. CD4 count in deciding whether to start anti-retroviral treatment in human immunodeficiency virus infection) with a strict cut-off, above or below which an intervention is assigned, and assumes that there will be little difference in subjects marginally over or under the asymptotic cut-off.[98] Assignment to intervention cannot be caused by the intervention but does require all participants to belong to the same population. The groups are then compared by observing the differences in subjects close to the asymptotic cut-off. The effect is measured by discontinuity from regression, which has been demonstrated mathematically to yield an unbiased estimate of a causal effect.[98]

Propensity scores: aim to provide less biased estimates of treatment effect and can be used for matching exposed and unexposed participants in a *case-control study* or to exclude non-overlapping data from analysis on the basis of an understanding of co-variates that affect the condition being studied.[100] They do so by attempting to correct for the non-balanced distribution of characteristics between the exposed and unexposed groups and are more statistically efficient than multi-variable regression models traditionally used to control for known confounders in OP studies.[100] However, like multi-variable regression, propensity scores can only correct for known confounders rather than all confounders but unlike multi-variable regression the sensitivity of propensity scores for unknown confounders can be estimated and reported.[100] In order to develop an effective propensity score a thorough understanding of the co-variates (i.e. the biology) is necessary for them to be included in the model.[100]

Focussing on dose-response relationship: one of Bradford Hill's criteria for causality is the presence of a dose-response relationship, where dose can be considered in terms of single daily dose and also as cumulative dose exposure over time. Thus one might expect, for example, to see a larger treatment effect in those exposed to a bigger single dose of an intervention (lower

blood pressure with higher dose of antihypertensive) and more adverse events occurring with a larger cumulative dose (more lung cancer with greater cumulative cigarette exposure). OP cohort studies have focussed inferences on the cumulative dose-response effect, such as in demonstrating that pioglitazone is not associated with an increased risk of bladder cancer.[101] Using a two-time updated exposure term, one of ever-/never-exposure and another of cumulative exposure has been shown mathematically to remove the allocation bias from the cumulative exposure term and provide a more reliable treatment effect estimate based on the cumulative exposure term.[102]

Instrumental variable analysis: can be applied to non-randomised studies to control for unmeasured confounding and has been used in OP cohort studies.[103] The method is underpinned by an 'instrument', which is linked to the treatment, but not directly or indirectly linked to the outcome except via the treatment being investigated. The challenge with such an approach is the identification of an instrument that must meet the following assumptions. First, the instrument should affect treatment allocation. Second, it should be a feature that is randomly assigned. Third, it should be associated with the outcome only via the treatment.[103,104] A good example to illustrate this might be differences between hospital formularies. In this case, the treatment's accessibility depends on it being included in a particular hospital's formulary, satisfying the first assumption. Although patients are not randomly allocated to hospitals, it might be acceptable to assume that, in general, patients do not present to a hospital due to knowledge of its formulary, satisfying the second assumption. Finally, so long as the hospital's formulary is not associated with other practices, such as quality-of-care for instance, the instrument can be thought to affect outcome only via the treatment itself, satisfying the third assumption.[103,104]

This means that in theory, by making these assumptions and collecting data on the instrument, it is possible to make treatment effect estimates on outcome without having to adjust for confounders.[103,104]

Case cross-over design of case-control studies: is a within-subject study design (compare with xRCT) attractive to OP, albeit appropriate only in specific circumstances. A comparison is made between the event time-window and the control time-window in terms of exposure. The within-patient control acts to block the effect of unmeasured between-patient time-invariant confounders without the need for these to be measured and prevents selection bias (as users are compared to themselves). However, this type of study design requires certain assumptions to be met in order to give valid results. First, the exposure must be short-lived and the outcome acute. Second, the risk associated with the exposure must rise and fall rapidly. These assumptions mean that the investigation of chronic diseases with long-term therapy is unsuitable with this type of study. [105–107]

Although in theory a case-cross over design could be used to investigate treatment effect, it has more often been used to assess potential harms, such as demonstrating that recent vaccination does not appear to raise the risk of multiple sclerosis relapse.[108] Importantly, this type of study design cannot account for time-varying within-patient confounders, such as changes to body mass index or smoking status. It also cannot be deployed when rates of drug exposure change across the time period being investigated, by, for instance, a new drug with the same indication being released. The case-crossover design is also sensitive to misspecification of the exposure window (see risk window bias) and if the drug is available over-the-counter, non-prescribed doses would be omitted from the patient's prescribing record leading to information bias. This study design is also prone to recall bias, when patients' recollections are used to define exposure rather than more objective measures, such as prescriptions.[105–107]

Partial blinding: Partial blinding involves the blinding of some aspects of an OP study (or indeed an RCT), for example observer blinding; although most OP studies by their very nature do not utilise randomisation, it is still possible to employ some form of blinding. The preferred technique is to separate the extraction of exposure information from outcome information.[109]

795 The published report should explain who was blinded and who was not as this helps with
796 critical appraisal.[109]

797

798 ***Missing data***

799

800 Although missing data can occur in both RCTs and OP studies, RCTs often include protocols that
801 go to great lengths to reduce this phenomenon. The collection of complete data may prove more
802 challenging in OP studies, where data is collected through routine clinical practice. Missing data
803 can lead to biased estimates particularly if it is not missing at random. There are a number of
804 techniques detailed in table 8 to handle missing data, although it remains the case that the best
805 way to deal with missing data is to prevent it from happening in the first place.[110]

806

Some specific biases in OP

The subsequent paragraphs give details about some specific biases in OP to be aware of:

Protopathic bias: occurs when the prescription of a treatment is caused by the symptoms of an undiagnosed condition. An example would be pancreatic cancer causing diabetes, leading to the prescription of an anti-diabetic drug. It may then appear as if the drug had caused the pancreatic cancer, when in fact the cancer had caused the indication for the drug – it is a form of reverse causality.[111] This bias may be detected in sensitivity analysis by comparing lag times of differing length from the first date of exposure to the development of the outcome.[112]

Surveillance bias in OP studies: or detection bias, is a differential (non-random) information bias, where one group of patients is more likely to have the outcome (or symptom associated with the outcome) diagnosed because of increased surveillance, screening or testing for the outcome. An example might be the use of ultra-sound Doppler (USS-D) for the diagnosis of deep-vein thrombosis (DVT) following trauma. Centres routinely screening all trauma patients for DVT are likely to have a higher rate of DVT diagnosis (and consequently treatment) than centres employing a symptom-, or risk score-based approach to USS-D in trauma patients.[113] This bias can be reduced by employing an unexposed comparison group with a similar pre-test probability of being screened, using outcomes thought to be diagnosed equally between the groups or adjusting for the differential detection rate in the analysis.[29]

Time-related biases in OP studies

Immortal time bias: is often introduced into OP studies by the definition of exposure or by the subsequent analysis. This is an important misclassification bias, a type of information bias. It

refers to a period of follow-up time between cohort entry and first drug exposure when the outcome of interest could not have occurred. Misclassification of the *pre-exposure* person-time as exposed or simply not counting the pre-exposure person-time leads to this bias, where the effect estimate is mistakenly skewed towards the treatment group. This bias is remedied by ensuring the *pre-exposure* time is counted, classified and analysed as *unexposed* person-time.[114–116]

Confounding by disease stage: is another form of information bias and can occur when comparing first-line therapy with subsequent treatment options. Those on first-line treatment are likely to be at an earlier stage of their disease compared those and second- or third-line treatment. Thus an outcome related to first-line therapy (and more likely to be prescribed to those with shorter disease duration) might be misattributed to subsequent treatment (more likely to be prescribed to those with longer disease duration), especially is there is a long lag between exposure and outcome. This can be avoided by comparing treatments in patients with similar disease duration/stage.[117]

Risk window bias: is specific to case-control studies. When considering, say, adverse drug reactions, the risk window is the period following exposure when the risk of the outcome is in excess of the background risk. In practice, the risk window can be extremely challenging to define and if it is too large serves to under-estimate the risk of the adverse event. It is best handled by sensitivity analysis comparing varying risk window durations.[118]

Conclusions

Undoubtedly, OP studies and RCTs have both contributed significantly to the evidence informing clinical practice. However, there is room for improvement to both types of approach.

Inferences based on RCT data are more likely to identify causal associations. This is because RCT reduce biases and confounding, meaning the effects detected are more likely to be caused by the treatment. However, RCTs do have shortcomings in relation to their external validity and their ability to detect harms. Moreover, when deployed inappropriately, without an evidence-based hypothesis, if there is failure to follow the ITT principle or they report multiple unplanned *post-hoc* subgroup analyses their findings may be misleading.

OP studies can complement the findings of RCTs and extend their results. However, caution should be exercised in their interpretation since there is the risk that the results observed represent bias or confounding. This is especially the case when making causal inferences from a small or unexpected treatment effect. There is an urgent need to train clinicians to understand robust study design and data analysis methods in OP to better appraise which studies provide valid evidence and which do not.

The pre-publication of study protocols and subgroup analysis alongside the adherence to reporting guidelines (CONSORT and STROBE) improves quality and aids critical appraisal of both study type. Also, design improvements or new variants of RCTs and OP studies may provide methodological advantages and, for OP studies in particular, may improve our confidence in their results. Combining evidence from both types of study in a considered and balanced fashion would also benefit patients.

It remains the case that, all things being equal, RCTs provide *better quality evidence* than OP studies but the latter, when well-conducted, can provide evidence with considerable clinical utility that may not be provided by RCTs.

Conflict of Interest

There are no conflicts of interest to declare within the submitted work.

The following authors have disclosed declarations of interest outside the submitted work:

TMC is funded by Diabetes UK with support from the British Heart Foundation. JMD holds a grant from PledPharma AB. HMC received grants (as part of EU Innovative Medicines programme collaborations) from AstraZeneca LP, Boehringer Ingelheim, Eli Lilly & Company, Pfizer, Roche Pharmaceuticals and Sanofi Aventis, and grants from Novo Nordisk. HMC is a shareholder in Bayer and Roche Pharmaceuticals. HMC is on trial steering committees or safety monitoring committees with Eli Lilly, Sanofi and Regeneron, Novartis Pharmaceuticals and Novo Nordisk and receives remuneration via her institution for this. She has received speaker fees and travel expenses for presenting trials she has helped design or other research she has led from Pfizer, Eli Lilly, Sanofi and Regeneron. DJW is Vice Chair of MHRA but this paper is written in the capacity as a University of Edinburgh academic and the paper does not reflect the views of MHRA.

References

1. Doll R, Hill AB. Smoking and carcinoma of the lung. *BMJ*. 1950;2(4682):739-748.

2. Banting FG, Best CH, Collip JB, Campbell WR, Fletcher AA. Pancreatic extracts in the treatment of diabetes mellitus. *Can Med Assoc J*. 1922;12(3):141-146.

3. IMI GetReal - Real-Life Data In Drug Development > Home. http://www.imi-getreal.eu/. Accessed January 8, 2019.

4. Akobeng AK. Understanding randomised controlled trials. *Arch Dis Child*. 2005;90(8):840-844. doi:10.1136/adc.2004.058222

5. Singal AG, Higgins PDR, Waljee AK. A Primer on Effectiveness and Efficacy Trials. *Clin Transl Gastroenterol*. 2014;5(1):e45. doi:10.1038/ctg.2013.13

6. Sources of evidence for assessing the safety, efficacy and effectiveness of medicines | The Academy of Medical Sciences. https://acmedsci.ac.uk/policy/policy-projects/methods-of-evaluating-evidence. Accessed January 8, 2019.

7. Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators. *Clin Epidemiol*. 2017;9:331-338. doi:10.2147/CLEP.S129728

8. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58(5):295-300.

9. Torgerson DJ, Roberts C. Randomisation methods: concealment. *BMJ*. 1999;319(7206):375-376. doi:10.1136/bmj.319.7206.375

10. Gordis L. Assessing preventive and therapeutic measures: randomized trials, assessing preventive and therapeutic measures: randomized trials. In: *Epidemiology*. 5th ed.

Philadelphia, PA: Elsevier/Saunders; 2014:138-154. https://www-clinicalkey-com.ezproxy.is.ed.ac.uk/#!/content/book/3-s2.0-B9781455737338000072. Accessed August 1, 2019.

11. Shrier I, Boivin J-F, Steele RJ, et al. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am J Epidemiol*. 2007;166(10):1203-1209. doi:10.1093/aje/kwm189

12. Cartwright N. What are randomised controlled trials good for? *Philos Stud*. 2010;147(1):59-70. doi:10.1007/s11098-009-9450-2

13. Sedgwick P. What is an open label trial? *BMJ*. 2014;348:g3434. doi:10.1136/bmj.g3434

14. Day SJ, Altman DG. Blinding in clinical trials and other studies. *BMJ*. 2000;321(7259):504. doi:10.1136/bmj.321.7259.504

15. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. *Fam Med*. 2007;39(2):132-137.

16. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317(7167):1185-1190. doi:10.1136/bmj.317.7167.1185

17. Hansson L, Hedner T, Dahlöf B. Prospective Randomized Open Blinded End-point (PROBE) Study. A novel design for intervention trials. *Blood Press*. 1992;1(2):113-119. doi:10.3109/08037059209077502

18. A Word About Evidence: 6. Bias—a proposed definition. Catalog of Bias. https://catalogofbias.org/2018/06/15/a-word-about-evidence-6-bias-a-proposed-definition/. Published June 15, 2018. Accessed January 8, 2019.

19. Porta M. *Dictionary of Epidemiology*. Oxford, UNITED STATES: Oxford University Press, Incorporated; 2014. http://ebookcentral.proquest.com/lib/ed/detail.action?docID=1679277. Accessed January 8, 2019.

20. Sedgwick P. Randomised controlled trials: understanding confounding. *BMJ*. 2015;351:h5119.

21. Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ*. 2000;321(7256):255-256.

22. Cochrane Handbook for Systematic Reviews of Interventions. http://handbook.cochrane.org/. Accessed May 13, 2017.

23. Schulz KF. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*. 2010;152(11):726. doi:10.7326/0003-4819-152-11-201006010-00232

24. Turner L, Shamseer L, Altman DG, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. In: *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd; 2012. doi:10.1002/14651858.MR000030.pub2

25. Hopewell S, Dutton S, Yu L-M, Chan A-W, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ*. 2010;340(1):c723-c723. doi:10.1136/bmj.c723

26. Plint AC, Moher D, Morrison A, Schulz K, al et. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust Pyrmont*. 2006;185(5):263-267.

27. Statistical Principles for Clinical Trials : ICH. http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html. Accessed January 8, 2019.

28. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA*. 2017;318(23):2337-2343. doi:10.1001/jama.2017.18556

29. ENCePP Home Page. http://www.encepp.eu/standards_and_guidances/methodologicalGuide4.shtml. Accessed January 8, 2019.

30. Behi R, Nolan M. Causality and control: threats to internal validity. *Br J Nurs*. 1996;5(6):374-377. doi:10.12968/bjon.1996.5.6.374

31. Levin KA. Study design VII. Randomised controlled trials. *Evid Based Dent*. 2007;8(1):22-23. doi:10.1038/sj.ebd.6400473

32. Kim H, Gurrin L, Ademi Z, Liew D. Overview of methods for comparing the efficacies of drugs in the absence of head-to-head clinical trial data. *Br J Clin Pharmacol*. 2014;77(1):116-121. doi:10.1111/bcp.12150

33. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet*. 2005;365(9453):82-93. doi:10.1016/S0140-6736(04)17670-8

34. Dumville JC, Torgerson DJ, Hewitt CE. Research methods: reporting attrition in randomised controlled trials. *BMJ*. 2006;332(7547):969-971.

35. Hotopf M. The pragmatic randomised controlled trial. *Adv Psychiatr Treat*. 2002;8(5):326-333. doi:10.1192/apt.8.5.326

36. Vestbo J, Leather D, Diar Bakerly N, et al. Effectiveness of Fluticasone Furoate-Vilanterol for COPD in Clinical Practice. *NEJM*. 2016;375(13):1253-1260. doi:10.1056/NEJMoa1608033

37. Strom BL, Eng SM, Faich G, et al. Comparative Mortality Associated With Ziprasidone and Olanzapine in Real-World Use Among 18,154 Patients With Schizophrenia: The Ziprasidone Observational Study of Cardiac Outcomes (ZODIAC). *Am J Psychiatry*. 2011;168(2):193-201. doi:10.1176/appi.ajp.2010.08040484

38. Fröbert O, Lagerqvist B, Olivecrona GK, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. *NEJM*. 2013;369:1587-1597. doi:DOI: 10.1056/NEJMoa1308789

39. Abraham J. International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use. In: Brouder A, Tietje C, eds. *Handbook of Transnational Economic Governance Regimes*. Brill; 2009:1041-1054. doi:10.1163/ej.9789004163300.i-1081.897

40. MRC/Wellcome Trust Workshop: regulation and biomedical research. 14/05 2008. https://wellcome.ac.uk/sites/default/files/wtx053463_0.pdf.

41. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312(7040):1215-1218.

42. Toussaint B. EU Clinical Trials Regulation. *Lancet*. 2013;381(9879):1719-1720. doi:10.1016/S0140-6736(13)61078-8

43. Donner A, Klar N. Pitfalls of and Controversies in Cluster Randomization Trials. *Am J Public Health*. 2004;94(3):416-422.

44. Khanna R, Bressler B, Levesque BG, et al. Early combined immunosuppression for the management of Crohn's disease (REACT): a cluster randomised controlled trial. *The Lancet*. 2015;386(10006):1825-1834. doi:10.1016/S0140-6736(15)00068-9

45. Joffe S. Evaluating novel therapies during the Ebola epidemic. *JAMA*. 2014;312(13):1299-1300. doi:10.1001/jama.2014.12867

46. Henao-Restrepo AM, Camacho A, Longini IM, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *Lancet*. 2017;389(10068):505-518. doi:10.1016/S0140-6736(16)32621-6

47. Camacho A on behalf of, Ebola ça suffit ring vaccination trial Consortium. The ring vaccination trial: a novel cluster randomised controlled trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola. *BMJ*. 2015;351:h3740. doi:10.1136/bmj.h3740

48. Sibbald B, Roberts C. Understanding controlled trials Crossover trials. *BMJ*. 1998;316(7146):1719-1720.

49. Sedgwick P. What is a crossover trial? *BMJ*. 2014;348:g3191. doi:10.1136/bmj.g3191

50. Dwosh IL, Giles AR, Ford PM, Pater JL, Anastassiades TP. Plasmapheresis Therapy in Rheumatoid Arthritis. *N Engl J Med*. 1983;308(19):1124-1129. doi:10.1056/NEJM198305123081903

51. Torgerson DJ, Torgerson C. Factorial RCTs. In: *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*. London, UNITED KINGDOM: Palgrave Macmillan Limited; 2008. http://ebookcentral.proquest.com/lib/ed/detail.action?docID=416816. Accessed August 1, 2019.

52. Sedgwick P. What is a factorial study design? *BMJ*. 2014;349:g5455. doi:10.1136/bmj.g5455

53. Bateman DN, Dear JW, Thanacoody HKR, et al. Reduction of adverse effects from intravenous acetylcysteine treatment for paracetamol poisoning: a randomised controlled trial. *The Lancet*. 2014;383(9918):697-704. doi:10.1016/S0140-6736(13)62062-0

54. Kovesdy CP, Kalantar-Zadeh K. Observational studies versus randomized controlled trials: avenues to causal inference in nephrology. *Adv Chronic Kidney Dis*. 2012;19(1):11-18. doi:10.1053/j.ackd.2011.09.004

55. Lawler M, Kaplan R, Wilson RH, Maughan T, on behalf of the S-CORT Consortium. Changing the Paradigm—Multistage Multiarm Randomized Trials and Stratified Cancer Medicine. *Oncologist*. 2015;20(8):849-851. doi:10.1634/theoncologist.2015-0014

56. Parmar MKB, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet*. 2014;384(9940):283-284. doi:10.1016/S0140-6736(14)61122-3

57. Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. *Ann Oncol*. 2017;28(1):34-43. doi:10.1093/annonc/mdw413

58. The STAMPEDE Trial. STAMPEDE. http://www.stampedetrial.org/. Accessed January 8, 2019.

59. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Stat Med*. 2014;33(19):3269-3279. doi:10.1002/sim.6183

60. Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *NEJM*. 2017;377(7):644-657. doi:10.1056/NEJMoa1611925

1050 61. Kasenda B, Schandelmaier S, Sun X, et al. Subgroup analyses in randomised controlled
1051 trials: cohort study on trial protocols and journal publications. *BMJ*. 2014;349:g4539.
1052 doi:10.1136/bmj.g4539

1053 62. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised
1054 controlled trials: systematic review. *BMJ*. 2012;344(mar15 1):e1553-e1553.
1055 doi:10.1136/bmj.e1553

1056 63. Rutter M, Pickles A. Annual Research Review: Threats to the validity of child psychiatry
1057 and psychology. *J Child Psychol Psychiatry*. 2016;57(3):398-416. doi:10.1111/jcpp.12461

1058 64. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility
1059 in data collection and analysis allows presenting anything as significant. *Psychol Sci*.
1060 2011;22(11):1359-1366. doi:10.1177/0956797611417632

1061 65. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic
1062 interventions. *Lancet*. 2008;372(9656):2152-2161. doi:10.1016/S0140-6736(08)61930-
1063 3

1064 66. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment
1065 effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266(1):93-98.
1066 doi:10.1001/jama.1991.03470010097038

1067 67. Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin*
1068 *Trials*. 2009;6(2):119-125. doi:10.1177/1740774509102310

1069 68. Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early.
1070 *BMJ*. 2012;344:e3863.

1071 69. Skovlund E. Repeated significance tests on accumulating survival data. *J Clin Epidemiol*.
1072 1999;52(11):1083-1088. doi:10.1016/S0895-4356(99)00090-6

70. Snapinn S, Chen M-G, Jiang Q, Koutsoukos T. Assessment of futility in clinical trials. *Pharm Stat*. 2006;5(4):273-281. doi:10.1002/pst.216

71. Pocock SJ. When To Stop A Clinical Trial. *BMJ*. 1992;305(6847):235-240.

72. Pocock SJ. Current controversies in data monitoring for clinical trials. *Clin Trials J Soc Clin Trials*. 2006;3(6):513-521. doi:10.1177/1740774506073467

73. Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions. In: *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008. http://www.ncbi.nlm.nih.gov/books/NBK47098/. Accessed August 1, 2019.

74. Rosenbaum PR. Observational Studies. In: *Observational Studies*. New York, NY: Springer New York; 2002:1-17. doi:10.1007/978-1-4757-3692-2_1

75. Medical Research Council MRC. Health Data Research UK (HDR UK). https://www.mrc.ac.uk/about/institutes-units-centres/uk-institute-for-health-and-biomedical-informatics-research/. Published June 1, 2017. Accessed January 8, 2019.

76. Kelly CM, Juurlink DN, Gomes T, et al. Selective serotonin reuptake inhibitors and breast cancer mortality in women receiving tamoxifen: a population based cohort study. *BMJ*. 2010;340:c693. doi:10.1136/bmj.c693

77. Steckler A, McLeroy KR. The importance of external validity. *Am J Public Health*. 2008;98(1):9-10. doi:10.2105/AJPH.2007.126847

78. Silverman SL. From randomized controlled trials to observational studies. *Am J Med*. 2009;122(2):114-120. doi:10.1016/j.amjmed.2008.09.030

79. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-349. doi:10.1016/j.jclinepi.2007.11.008

80. Costa BR da, Cevallos M, Altman DG, Rutjes AWS, Egger M. Uses and misuses of the STROBE statement: bibliographic study. *BMJ Open*. 2011;1(1):e000048. doi:10.1136/bmjopen-2010-000048

81. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*. 2014;383(9921):999-1008. doi:10.1016/S0140-6736(13)61752-3

82. Mauro MJ, Davis C, Zyczynski T, Khoury HJ. The role of observational studies in optimizing the clinical management of chronic myeloid leukemia. *Ther Adv Hematol*. 2015;6(1):3-14. doi:10.1177/2040620714560305

83. Berlin JA, Glasser SC, Ellenberg SS. Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am J Public Health*. 2008;98(8):1366-1371. doi:10.2105/AJPH.2007.124537

84. Chan EW, Liu KQL, Chui CSL, Sing C-W, Wong LYL, Wong ICK. Adverse drug reactions – examples of detection of rare events using databases. *Br J Clin Pharmacol*. 2015;80(4):855-861. doi:10.1111/bcp.12474

85. Struck R, Baumgarten G, Wittmann M. Cost-efficiency of knowledge creation: randomized controlled trials vs. observational studies. *Curr Opin Anaesthesiol*. 2014;27(2):190-194. doi:10.1097/ACO.0000000000000060

86. Clinical Practice Research Datalink - CPRD. https://www.cprd.com/intro.asp. Accessed January 8, 2019.

87. SCI-Diabetes. http://www.sci-diabetes.scot.nhs.uk/. Accessed January 8, 2019.

88. Health Maintenance Organization Research Network (HMORN) UCSF Center for Diabetes Translational Research | Global Research Projects. https://globalprojects.ucsf.edu/project/health-maintenance-organization-research-network-hmorn-ucsf-center-diabetes-translational. Accessed January 8, 2019.

89. Chow JTY, Lam K, Naeem A, Akanda ZZ, Si FF, Hodge W. The pathway to RCTs: how many roads are there? Examining the homogeneity of RCT justification. *Trials*. 2017;18. doi:10.1186/s13063-017-1804-z

90. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359(9302):248.

91. Ranstam J. Bias in observational studies. *Acta Radiol*. 2008;49(6):644-645. doi:10.1080/02841850802075082

92. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *NEJM*. 2000;342(25):1907-1909. doi:10.1056/NEJM200006223422511

93. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *NEJM*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507

94. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *NEJM*. 2000;342(25):1878-1886. doi:10.1056/NEJM200006223422506

95. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. In: The Cochrane Collaboration, ed. *Cochrane Database of Systematic Reviews*. Chichester, UK: John Wiley & Sons, Ltd; 2014. doi:10.1002/14651858.MR000034.pub2

96. Gibbons CL, Mangen M-JJ, Plass D, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*. 2014;14:147. doi:10.1186/1471-2458-14-147

97. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. *Pharmacoepidemiol Drug Saf*. 2013;22(1):1-6. doi:10.1002/pds.3334

98. Rutter M. Proceeding from observed correlation to causal inference: the use of natural experiments. *Perspect Psychol Sci*. 2007;2(4):377-395. doi:10.1111/j.1745-6916.2007.00050.x

99. Uchiyama T, Kurosawa M, Inaba Y. MMR-Vaccine and Regression in Autism Spectrum Disorders: Negative Results Presented from Japan. *J Autism Dev Disord*. 2007;37(2):210-217. doi:10.1007/s10803-006-0157-3

100. Okoli GN, Sanders RD, Myles P. Demystifying propensity scores. *Br J Anaesth*. 2014;112(1):13-15. doi:10.1093/bja/aet290

101. Levin D, Bell S, Sund R, et al. Pioglitazone and bladder cancer risk: a multipopulation pooled, cumulative exposure analysis. *Diabetologia*. 2015;58(3):493-504. doi:10.1007/s00125-014-3456-9

102. Farran B, McGurnaghan S, Looker HC, et al. Modelling cumulative exposure for inference about drug effects in observational studies. *Pharmacoepidemiol Drug Saf*. 2017;26(12):1527-1533. doi:10.1002/pds.4327

103. Hernán MA, Robins JM. Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology*. 2006;17(4):360. doi:10.1097/01.ede.0000222409.00878.37

104. Jamal Uddin M KO. Instrumental Variable Analysis in Epidemiologic Studies: An Overview of the Estimation Methods. *Pharm Anal Acta*. 2015;06(04). doi:10.4172/2153-2435.1000353

105. Etminan M, Samii A. Pharmacoepidemiology I: A Review of Pharmacoepidemiologic Study Designs. *Pharmacotherapy*. 2004;24(8):964-969. doi:10.1592/phco.24.11.964.36143

106. "Chris" Delaney JA, Suissa S. The case-crossover study design in pharmacoepidemiology. *Stat Methods Med Res*. 2009;18(1):53-65. doi:10.1177/0962280208092346

107. Donnan PT, Wang J. The case-crossover and case-time-control designs in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*. 2001;10(3):259-262. doi:10.1002/pds.590

108. Confavreux C, Suissa S, Saddier P, Bourdès V, Vukusic S. Vaccinations and the Risk of Relapse in Multiple Sclerosis. *N Engl J Med*. 2001;344(5):319-326. doi:10.1056/NEJM200102013440501

109. Parker RA, Berman NG. Chapter 28 - Blinding in observational studies. In: *Planning Clinical Research*. Cambridge: Cambridge University Press; 2016. doi:10.1017/CBO9781139024716

110. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402

111. Korhonen MJ, Huupponen R, Ruokoniemi P, Helin-Salmivaara A. Protopathic bias in observational studies on statin effectiveness. *Eur J Clin Pharmacol*. 2009;65(11):1167. doi:10.1007/s00228-009-0701-0

1184 112. Potential overestimation of risk by protopathic bias and mitigation by the introduction of
1185 lag-time. October 2018. https://www.bmj.com/content/354/bmj.i4857/rr-3. Accessed
1186 August 1, 2019.

1187 113. Haut ER, Pronovost PJ. Surveillance Bias in Outcomes Reporting. *JAMA*.
1188 2011;305(23):2462-2463. doi:10.1001/jama.2011.822

1189 114. Agarwal P, Moshier E, Ru M, et al. Immortal Time Bias in Observational Studies of Time-
1190 to-Event Outcomes. *Cancer Control J Moffitt Cancer Cent*. 2018;36(1):195-199.
1191 doi:10.1177/1073274818789355

1192 115. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort
1193 studies: example using statins for preventing progression of diabetes. *BMJ*.
1194 2010;340(7752):907-911.

1195 116. Suissa S. Immortal Time Bias in Pharmacoepidemiology. *Am J Epidemiol*.
1196 2008;167(4):492-499. doi:10.1093/aje/kwm324

1197 117. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational
1198 studies. *Diabetes Care*. 2012;35(12):2665-2673. doi:10.2337/dc12-0788

1199 118. van Staa TP, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification
1200 due to the time-window design in pharmacoepidemiologic studies. *J Clin Epidemiol*.
1201 1994;47(2):183-189.

1202 119. Glossary. NICE. https://www.nice.org.uk/glossary?letter=a. Accessed January 8, 2019.

1203 120. Sedgwick P. Selection bias versus allocation bias. *BMJ*. 2013;346:f3345.
1204 doi:10.1136/bmj.f3345

121. Viswanathan M, Berkman ND, Dryden DM, Hartling L. *Approaches to Assessing the Risk of Bias in Studies*. Agency for Healthcare Research and Quality (US); 2013. https://www.ncbi.nlm.nih.gov/books/NBK154465/. Accessed April 9, 2018.

122. Hammer GP, Prel J-B du, Blettner M. Avoiding bias in observational studies: part 8 in a series of articles on evaluation of scientific publications. *Dtsch Ärztebl Int*. 2009;106(41):664. doi:10.3238/arztebl.2009.0664

123. Sheehan NA, Didelez V, Burton PR, Tobin MD. Mendelian randomisation and causal inference in observational epidemiology. *PLOS Med*. 2008;5(8):e177. doi:10.1371/journal.pmed.0050177

124. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence?study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407-415. doi:10.1016/j.jclinepi.2010.07.017

125. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence?inconsistency. *J Clin Epidemiol*. 2011;64(12):1294-1302. doi:10.1016/j.jclinepi.2011.03.017

126. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence?indirectness. *J Clin Epidemiol*. 2011;64(12):1303-1310. doi:10.1016/j.jclinepi.2011.04.014

127. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence?imprecision. *J Clin Epidemiol*. 2011;64(12):1283-1293. doi:10.1016/j.jclinepi.2011.01.012

128. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence?publication bias. *J Clin Epidemiol*. 2011;64(12):1277-1282. doi:10.1016/j.jclinepi.2011.01.011

129. Guyatt GH, Oxman AD, Vist GE, et al. Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.

130. Heneghan C. Rare adverse events in clinical trials: understanding the rule of three. BMJ EBM Spotlight. http://blogs.bmj.com/bmjebmspotlight/2017/11/14/rare-adverse-events-clinical-trials-understanding-rule-three/. Published November 14, 2017. Accessed January 8, 2019.

131. Tavazzi L. Do we need clinical registries? *Eur Heart J*. 2014;35(1):7-9. doi:10.1093/eurheartj/eht360

1239

Randomised control trial

A study in which a number of similar people are randomly assigned to two (or more) groups to test an intervention. One group (or more) has the intervention and others act as a control (alternative intervention, placebo or no interventional at all). Outcomes are measured at specific times and any difference in response is measured statistically.

Observational Study

A prospective or retrospective study in which the investigator observes the natural course of events, with or without a control group. Rather than being randomly assigned, the intervention is chosen for, or by, the patient. Any difference in results is measured statistically.

Table 1: Brief description of main study types (modified from NICE)[119]

1240

1241

Bias	Mitigation in RCTs	Mitigation in observational studies
Selection bias: occurs where individuals are more likely to be selected for a study than others, meaning that the patients included in the study are different from those who are not. [120]	Randomising all patients who are eligible for inclusion who consent Flow chart (as recommended in CONSORT[23]) demonstrating patient characteristics not included in study, including those who don't consent Blinding	Flow chart demonstrating patient characteristics not included in study, including those who don't consent
Allocation bias: occurs due to absence of comparability between groups in the allocation of treatment such that they differ significantly from one another by a factor other than the disease or exposure under investigation. [90,91]	Randomisation Blinding	Matching for a control group

Information bias[1]: occurs when information is obtained differently between exposed and unexposed cases such as a flaw in measuring exposure, outcomes or co-variates with differing accuracy between groups. For continuous variables this is known as measurement error, for discrete variables classification error. [90,91]	Blinding Using standardised, validated tools to collect data	Blinding of outcome/exposure Using standardised, validated tools to collect data
Performance bias: difference between the groups in the way that care is provided or subjects are exposed to factors other than the intervention under study.[121,122]	Blinding Reporting unblinding, where this becomes necessary	Blinding of exposure

[1] Differential information bias tends to exaggerate an association in either direction, where the bias functions to change the likelihood of exposed or unexposed cases being identified such that one or the other is *unequally* likely to be identified and recorded. Non-differential information bias, exposed and unexposed cases are affected *equally*, where all data might be gathered through an unreliable measure and thus test power is reduced and the association tends to be under-estimated.

Confounding: refers to the mixture or blurring of effects. This often occurs during the statistical analysis of data gathered in observational studies, where an association is found between an exposure and an outcome but the association found is, in fact, due the observation of the effect of another, unmeasured factor. [90,91,123]	Randomisation Blinding Stratified randomisation	Multi-variable regression analysis, for known confounders
Attrition bias: unequal loss of participants between the groups such that they are no longer similar to one another. [121,122]	ITT analysis Flow chart demonstrating patient characteristics lost to follow-up	ITT analysis Flow chart demonstrating patient characteristics lost to follow-up
Publication bias: when publication depends on the hypothesis being tested and the significance and direction of the effects detected. [121,122]	Pre-publishing study protocols Publishing study regardless of outcome	Pre-publishing study protocols Publishing study regardless of outcome

Not identifying unplanned, *post-hoc* sub-group analysis or, if done, not labelling them as hypothesis-generating analyses	Not identifying unplanned, *post-hoc* sub-group analysis or, if done, not labelling them as hypothesis-generating analyses

Table 2: Major biases and how they can be mitigated in RCTs and observational studies; see https://catalogofbias.org/biases/ for further detail about bias

1242

Strength of the association	The stronger the association, the more likely the effect is causal
Consistency	Reproducibility
Specificity	A specific exposure gives rise to a specific outcome
Temporality	The exposure must precede the outcome
Biological gradient	A dose-response relationship; the greater the exposure the larger the effect
Plausibility	Consistent with scientific understanding
Coherence	Coherent with other theories
Experiment	The outcome can be altered, improved or abolished by experiment – '*here the strongest support for causation can be revealed*'[8]

Table 3: Bradford Hill's criteria for causality as applied to medical sciences [8]

1243

1244

Study limitations (risk of bias) [124]	Failure to conceal allocation
	Failure to blind
	Loss to follow-up
	Failure to consider intent-to-treat principle
	Stopping early for benefit
	Use of unvalidated outcome measures
	Carry-over effects in cross-over trials
	Recruitment bias in cluster-randomised trials (if those recruiting participants know the participants' allocation, even when allocation of clusters has been adequately concealed)
Inconsistency of results [125]	Point estimates vary widely across studies
	Confidence intervals show minimal or no overlap
	The statistical test for heterogeneity shows a low *p*-value
	I^2, a statistical test for heterogeneity, is large
Indirectness of evidence [126]	Differences in populations
	Differences in interventions
	Differences in outcomes
	Indirect comparisons
Imprecision (random error) [127]	Insufficient sample size
	Low event rate
	Confidence interval overlaps no effect

Reporting bias [124,128]	Publication bias: consider especially when only a small number of commercially funded trials available Selective reporting bias: consider when there is non-publication of original study protocol

Table 4: Factors which might reduce confidence in an RCT, either when considered alone or when compared in meta-analysis [129]

1245

	Examples
Setting of the trial	Country
	Healthcare system/standard of care
Selection of patients	Inclusion/exclusion criteria
	Run-in periods, where non-tolerant patients are excluded
Characteristics of randomised patients	Baseline characteristics
	Underlying homogeneity of pathology
	Presence of comorbidities
Differences between trial protocol and routine practice	Prohibition of non-trial treatments
	Timing of treatment in disease course
Outcome measures and follow-up	Surrogate outcome measures

	Use of patient-centred measures
	Relevance of composite endpoints
Adverse effects of treatment	Rates of discontinuation
	Completeness of reporting
	Exclusion of patients at risk of adverse outcomes

Table 5: Factors to consider when assessing generalisability [33]

1246

Expected incidence of adverse drug reaction	Number of patients to be observed to detect:		
	1 event	**2 events**	**3 events**
1:100	300	480	650
1:200	600	960	1,300
1:1,000	3,000	4,800	6,500
1:2,000	6,000	9,600	13,000
1:10,000	30,000	48,000	65,000

Table 6: Number of patients to be observed to detect a given adverse event, modified from [130]

Prospective evaluation of patient population and disease characteristics

Assessment and comparison of costs and effectiveness associated with diagnostics

Investigation of adherence to guidelines

Post-marketing surveillance

Detection of responsive sub-groups

Characterisation of risk factors and levels of risk

Identification of relevant sources of uncertainty

Cost evaluation

Formation of hypotheses to be tested in subsequent experiments

Table 7: Particular areas suited to the use of observational studies[131]

1251

1252

Method	Description
Listwise/case deletion	• Simply omits the subjects in whom the data are absent. If the missing data occur randomly, then this method produces unbiased results. However, data points are often not missing at random, and in this case listwise deletion will lead to biased estimates of treatment effect.
Pairwise deletion	• Omits information only when data testing a particular assumption are missing; if it is missing from elsewhere existing values are used instead. This may lead to modelling problems where sample sizes and standard errors of co-variates differ from one another.
Mean substitution	• The missing value of a variable is replaced by its mean value from other subjects. This method gains no new information (as it is created from information that exists already), and leads to bias when the data are not missing at random. This is generally not an accepted approach.
Regression imputation	• Uses regression modelling to estimate missing values, but like mean substitution adds to no information.
Last observation carried forward	• Replaces absent data with the last recorded value for all missing data points. Although this approach is simple, it under-estimates intra-subject variability and gives rise to an illusion of precision.
Maximum likelihood modelling	• Assumes that the data present all arise from a multi-variate normal distribution. If there are few missing data, the absent data points can be estimated by using the conditional distribution of other variables.

Expectation maximisation	• Utilises maximum likelihood modelling to create an entirely new (modelled) dataset based on all the available information. The process is iterative and stops when the new dataset is stable. This approach is computer intensive, especially if there are many missing data, and tends to underestimate standard errors and thus overestimate precision.
Multiple imputation	• Replaces missing data with a range of plausible values representing the natural variability of values. A model is run, substituting each value in the range for each missing data point and a standard statistical analysis is on each iteration. Summary statistics are created by combining the statistic from each model run and is robust as it retains the variability and uncertainty of the missing data.
Sensitivity analysis	• Is an analysis that aims to characterise how uncertainty in the output can be attributed to uncertainty in the input. All methods dealing with missing data should be subjected to this form of analysis, by comparing effect estimates with and without these missing data and then to the method used to handle the missing data.

1253 ***Table 8: Examples of various methods employed to handle missing data, modified from*** [110]