Received: 22 January 2019 | Revised: 23 April 2019 | Accepted: 24 May 2019

DOI: 10.1111/bcp.14024

REVIEW

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

Thomas M. Caparrotta[1] | James W. Dear[2] | Helen M. Colhoun[1] | David J. Webb[2]

[1] Institute of Genetics and Molecular Medicine, University of Edinburgh, UK

[2] Queen's Medical Research Institute, University of Edinburgh, UK

Correspondence
Thomas Caparrotta, Clinical Pharmacology and Therapeutics University of Edinburgh, UK.
Email: tom.caparrotta@igmm.ed.ac.uk

Funding information
Diabetes UK, Grant/Award Number: 18/0005786

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 are the best way to determine whether an intervention modifies an outcome being studied, largely because randomisation reduces bias and confounding. Observational studies are useful to investigate whether benefits/harms of a treatment are seen in day-to-day clinical practice in a wider group of patients. 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.

KEYWORDS

clinical trial methodology, evidence-based medicine, health policy, pharmacoepidemiology, quality use of medicines

1 | INTRODUCTION

Robust evidence about clinical interventions is necessary for many reasons, from new treatment licensing to informing clinical practice, guideline creation and clinical/cost effectiveness analysis. Pharmacoepidemiology involves the study of drug-based interventions in populations and, for >70 years, the randomised controlled trial (RCT; see Glossary for all key terms) 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 treatment effect estimate (TEE) by determining whether an intervention modifies an outcome under study. For this reason, they are the preferred approach for estimating relative and absolute TEEs 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 also provide 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 are universally accepted for delineating the natural history of diseases, their risk factors and prognostic markers. However, OP, where (beneficial/harmful) *treatment effects* are quantified, has been subject to criticism because bias and confounding create difficulty in attributing cause and effect. Notwithstanding, OP studies are the mainstay of pharmacovigilance for harmful effects once a drug has been licensed. Indeed, following robust assessment of efficacy by RCTs, OP studies are able to assess whether an intervention is effective is day-to-day clinical practice, which often includes more heterogeneous patient groups and less precise diagnostic criteria than might feature in an RCT.

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 deliver evidence not supplied by the other. Indeed, 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*'.[3] Both RCTs and OP studies have strengths and weaknesses. Both provide flawed answers, through poor design, execution or analysis. There is also increasing concern about the observed efficacy–effectiveness gap and well-designed OP studies (alongside more generalisable RCTs) will help plug this.[4]

Robustly designed and conducted RCTs have good internal validity, allowing inferences on *efficacy/relative efficacy* and causality to be made.[5] *Relative effectiveness* can be measured in pragmatic RCTs or in OP studies.[4,6]

Evaluating sources of clinical evidence is a key skill for clinical decision-making. In light of this, we discuss the inherent properties, advantages and disadvantages of both types of study and how they might be improved to 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 conflict, the results of a well-designed and executed RCT are more likely to represent an unbiased TEE. 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.

2 | RANDOMISED CONTROLLED TRIALS IN PHARMACOEPIDEMIOLOGY

The inherent properties of RCTs make them the most robust means of evaluating healthcare interventions.[5] When properly designed and executed, with sufficient power and appropriate analysis, RCTs give the best indication of the *efficacy* of an intervention.[5] The *key properties* of RCTs that differ from OP studies are:

- A preplanned experiment, which gives rise to internal validity (and can reduce selection bias)
- Random treatment allocation, 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 (but not always done)
- Formal calculation of adequate trial size to ensure satisfactory study power and thus meaningful results
- Minimisation of missing data and systematic collection of outcomes to prevent information bias

Hypothesised effect modification can be measured in an RCT through stratification by the potential effect modifier's presence or absence, thus allowing the identification of people who may benefit from, or be harmed by, a given treatment.[7] If no effect modifiers affect a drug, its effects are said to be homogeneous. It is important that any such strategy be specified in advance.

Bradford Hill[8] lists several criteria that increase confidence that an association is causal (see Table 1 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 have limitations, assuming otherwise robust design. These relate particularly to the generalisability of results. Other limitations of RCTs include length of follow-up and trial size. When long and/or large, costs can increase dramatically and when inadequately so, 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

TABLE 1 Bradford Hill's criteria for causality as applied to medical sciences[8]

Strength of the association	The stronger the association, the more likely it is that 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]

of RCTs in pharmacoepidemiology and strategies to ensure their good conduct are addressed in more detail.

2.1 | The advantage of randomisation, allocation concealment and blinding

Randomisation, stratified or patient-level, is a major contributor to the benefit RCTs have over observational studies. Any increase in comparability between the groups caused by randomisation applies equally to variables we can and cannot measure as confounders and effect modifiers are reduced or balanced.[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. In RCTs, although blinding requires allocation concealment, allocation concealment is not always followed by blinding (open-label RCT).[13] Ideally, when a study *is* blinded, this should apply to all participants and staff, but is not always practicable (then called partial blinding).[13,14] Studies should always report who was blinded and who was not.

PROBE study: is a study type where outcome data are collected through routine clinical care and thought to increase the generalisability of findings. The open-label nature of the trial may introduce observer bias in the recording of the endpoints, even though the use of hard endpoints tries to reduce subjectivity. Also, patients know to which intervention they are exposed, introducing the risk of contamination if they seek treatment from another healthcare provider/over-the-counter that goes unrecorded in the trial.[15]

2.2 | Reduction of bias and confounding in RCTs compared to OP studies

The design of RCTs reduces bias and confounding and hence spurious and indirect associations respectively. There is not always a clear distinction between bias and confounding, but bias can be considered a *design flaw* and confounding a true, but not causal, association. For example, in OP because other factors (e.g. frailty) may be associated with both the propensity to be allocated a drug and with the outcome of interest, frailty is a confounder of the drug exposure–outcome association. Although this is sometimes referred to as confounding by indication, others refer to it as allocation bias since the allocation is non-random. In this case, however, the bias is not a study design effect.

In the main, there are 3 categories of bias which the design of an RCT minimises (indeed most biases fit into 1 of these broad categories, despite their varying nomenclature); selection bias, allocation bias and information bias.[16]

The Cochrane Handbook for Systematic Review of Interventions contains a tool that uses readers' judgement to assess the risk of bias in a study, and hence render a verdict about its internal validity and in turn whether it merits inclusion in evidence synthesis.[17] To maintain the benefits inherent to RCTs and provide for adequate reporting of protocols and results, the following paragraphs describe agreed reporting standards.

CONsolidated Standards Of Reporting Trials (CONSORT)[18]**:** since 1993, the reporting standard has undergone regular revisions; the current 2010 revision consists of a 25-item checklist and flowchart focussing on trial design, analysis and interpretation; a central tenet is the preregistration 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 preplanned or not and if not why. Studies have investigated the effect of the 2001 revision to the CONSORT guidelines on the completeness of reporting.[19-21] Although these found a general trend of improvement in the reporting of important aspects of trial methodology, it remained sub-optimal.[19-21] Endorsement of CONSORT by journals may beneficially influence the completeness of trial reporting.[19]

Statistical analysis plan: *(SAP)* 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 heavily influence a trials' conclusions.[22] Until 2017, no guidance for SAP contents existed (compare with CONSORT, around since 1993). A recently published expert consensus document has now specified minimum content for an SAP in relation to RCTs[23] (and now also for OP studies[24]). It will be important to measure whether this improves transparency of reporting statistical analysis and consequently whether this improves the reproducibility of RCTs (and OP studies).

2.3 | RCTs have internal validity, which allows causality to be established

Well-designed and well-conducted RCTs have internal validity.[25] This is especially the case when the population being examined is large and, by analogy, if the findings are replicated elsewhere. However, an RCT may produce a TEE not generalisable beyond the population being studied, despite having internal validity. Conversely, a generalisable RCT has a TEE which can be applied more broadly. However, in order to calculate an absolute risk reduction, the TEE obtained must apply to the background population against whom the absolute risk reduction is to be calculated.

2.4 | RCTs facilitate the comparison of treatments

Due to the results probably 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 noninferiority or superiority, comparing multiple treatments or dosages simultaneously, and are becoming more common.[26]

With an ever-increasing number of treatments available, it is impractical to carry out head-to-head comparisons of each one. A well-conducted RCT— due to the confidence that the TEE observed

is likely to be true — can more easily be incorporated into adjusted indirect comparisons, mixed treatment comparisons, meta-analyses and systematic reviews than OP studies. Thus, RCTs allow statistical inference to be made regarding the efficacy of different interventions even when direct comparison has not been made.[27]

3 | MAXIMISING THE RESULTS OF RCTS IN PHARMACOEPIDEMIOLOGY

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

Although preferable for assessment of efficacy, individual trial methodology must be scrutinised to critically appraise its results (see Table 2 for factors that might reduce confidence in RCTs or meta-analyses combining RCTs).

It is important to consider the context in which a trial has been conducted. For example, generalisability may be particularly compromised in pharmaceutical company-funded studies where the objective is to demonstrate superiority of a new drug over existing therapy and a highly selective study population has been used or if there is differential drop out between arms.

TABLE 2 Factors that might reduce confidence in a randomised controlled trial, either when considered alone or when compared in meta-analysis[28]

Study limitations (risk of bias)[29]	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[30]	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[31]	Differences in populations Differences in interventions Differences in outcomes Indirect comparisons
Imprecision (random error)[32]	Insufficient sample size Low event rate Confidence interval overlaps no effect
Reporting bias[29,33]	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

The *New England Journal of Medicine* has published an excellent series on RCTs, written by clinical trialists for clinical trialists, covering much of the material detailed below in more depth.[34]

3.1 | The findings of RCTs may not be generalisable

RCTs are often done 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 of patients receiving the intervention in the community. While these trials are critical to establish efficacy and preliminary safety, it may mean that the study results are only valid in the specific group of participants included in the trial (i.e. not generalisable). Also, some RCTs have been criticised for not taking into consideration factors important to patients' well-being.[35]

Generalisability may not apply to an RCT unless steps are taken to make it more generalisable, such as by using some of the following methods.[35]

Intention to treat analysis: (ITT) can help prevent attrition bias, which threatens the internal validity of RCTs by removing the benefits of randomisation, introducing potential bias, confounding and imbalance in the prior risk of the outcome of interest between study groups (and also reduce the study's power).[35,36] ITT evaluation might also better reflect real-world clinical practice (*increasing generalisability*), where patients may adhere poorly to, or stop, the intervention and thus gives a more realistic TEE (as these nonadherent patients are accounted for), although underestimating the maximum achievable benefit.[10,16]

Sensitivity analyses should be presented comparing relative TEE of per protocol results (those who fully completed the study protocol) to ITT results. It is also good practice to provide the subgroup characteristics of patients lost to follow-up.[36]

Pragmatic RCTs: (pRCTs) aim to redress perceived problems in generalisability of RCTs by providing answers to questions relevant to patients and clinicians.[37] The Salford Lung Study randomised ~50% of the community patient population, demonstrating the superiority of fluticasone furoate and vilanterol over usual care in the management of COPD, the results of which are broadly generalisable.[38]

Large simple RCTs: (lsRCTs) 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 as outcomes are measured from routine care, increasing generalisability. 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.[39]

Randomised database studies: are likely to become more common with the increasing availability of computerised health data e.g. a Swedish study demonstrating that thrombus aspiration prior to stenting in acute ST-elevation myocardial infarction was no better than stenting alone, with similar outcomes in all subgroups.[40]

3.2 | The role of meta-analysis in assessing generalisability

Meta-analyses involve combining the results of many RCTs to get a more precise estimate of the true TEE since the effective sample size is increased, increasing generalisability. An important aspect of meta-analysis is to test whether it is valid to combine data in this way or whether substantial unexplained heterogeneity (as measured by Cochran's C test or the I^2 statistic) renders the combined estimate invalid.[41]

3.3 | RCTs can be expensive and difficult to undertake

RCTs can be expensive to conduct. For instance, 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.[42,43] This is in part due to standards required by the study protocol to ensure internal validity, safety, regulation compliance and length of time required for follow-up.[10] Costs can increase further if larger groups of patients are needed to power RCTs.[10] Also, the outcome of interest may be so far in the future that it is difficult and expensive to maintain follow-up.[44]

Cluster RCTs: (cRCTs) have less statistical efficiency than randomising an equivalent number of people at the individual-level.[45] An example of a cRCT is the Randomised Evaluation of an Algorithm for Crohn's Treatment Trial that randomised gastroenterology clinics to standard incremental therapies for disease control or early combined immunosuppression and demonstrating no difference in primary outcome between the units of randomisation.[46]

It may be impossible to do an RCT in emergency situations; for example following a terrorist incident or during an epidemic where there is a need to produce information quickly while at the same time minimising the risk to patients and staff.[47] The cRCT can help with these difficulties. The 2014–2015 Ebola outbreak led to the design of novel approaches to undertaking RCTs. The adaptive ring vaccination, open-label, cRCT (Ebola *Ça Suffit* Trial) was used to demonstrate efficacy of rVSV-vectored Ebola vaccine, where immediate *vs.* delayed vaccination was compared and immediate vaccination was favoured.[48] 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.[49]

Crossover RCTs: (xRCTs) reduce intersubject variability (thus increasing precision) but are not appropriate if there is a significant carryover from one of the treatments, despite washout.[50,51] xRCTs can increase study power but cannot be used for conditions with an acute natural history nor investigate treatments providing cure rather than respite.[50] For example, an xRCT investigated sequential plasmapheresis *vs.* sham plasmapheresis (placebo) in the same patients to measure symptom improvement in rheumatoid arthritis and showed no difference between the treatments.[52]

Factorial RCT: (fRCT) is another design assessing outcomes more efficiently than separate trials.[53] An fRCT was used by investigators to gauge 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 if pretreatment with ondansetron reduced nausea and vomiting due to NAC treatment. The study found that both the shortened course of NAC and ondansetron pretreatment reduced nausea and vomiting independently and also additively (and that the shortened course of NAC was associated with fewer anaphylactoid reactions).[54]

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

Careful thought is given to planning RCTs to ensure internal validity and adequate power. Should insufficient participants be recruited, more participants than expected drop-out or a lower than expected event rate be observed then the trial power may be inadequate to detect significant change (or have to continue for longer than planned) to yield an adequate TEE.[55] 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 better than 2-arm studies at demonstrating superiority, which frequently do not show this.[56] Patients, clinicians and regulatory authorities want to know whether certain interventions beat those already available as quickly as possible.[56,57] The ideal study design is yet to be established and the multiple arms may cause difficulties interpreting results, particularly if arms are added/removed.

Adaptive clinical trials: (ACTs), with multiple arms, seek to address some of the concerns about multi-arms studies. To date, ACTs have been mainly deployed in the field of oncology, since their design can handle the increasingly-recognised biological heterogeneity of tumours but they show promise in other fields. There are different forms of these trials, but all allow some prespecified adaptation to take into account evolving understanding both from within and outside the trial. ACTs aim to address deficiencies in traditional trial design, described as the *weakest link* in cancer therapy development, given molecular understanding of tumour biology has increased.[58]

The STAMPEDE trial investigating treatments for advanced prostate cancer is one example of an adaptive, multi-arm, multi-stage platform trial.[59] In addition to the multiple cross-wise comparisons, the necessity to undertake repeated interim analyses make these trials complicated to analyse and interpret, as these increase the likelihood that a positive finding is significantly different by chance alone if not accounted for.[57,60] Types include: *basket trials*, *umbrella trials* and *platform trials*.

Stopping after a prespecified 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 their length of time under observation (although the

observed event rate is influenced by these 2 latter parameters). This may increase trial efficiency by declaring a TEE as significant on the basis of an observed difference in event rates between 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.[61]

3.5 | The advantages of prespecifying sub-group analysis

RCTs often report results with regards to sub-groups differing from each other in baseline traits. Trial populations are often heterogenous, raising the question of whether effects observed hold for all of the patients regardless of baseline characteristics.[62,63] Conducted appropriately, sub-group analysis is illuminating, increases generalisability and impacts positively on patient care.[62] However, performed poorly, or indeed not reported, it can be misleading.[62,63]

Preplanned 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.[18,62] However, systematic reviews have shown that reported sub-group analyses are seldom prespecified and there is a recognition that uncontrolled flexibility in the analysis of data carries a real risk of false positive findings.[62-65] If multiple assumptions are tested in subgroup analysis the likelihood of a falsely significant result by chance alone increases.[66] 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.[67]

3.6 | 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 TEE and denying potential users a beneficial new treatment. When a trial is stopped early its internal validity is compromised. A trial stopped early for beneficial reasons may *overestimate* the treatment effect because: the decision to stop trials early requires data analysis on multiple occasions; probability theory states, the more times data are analysed, the more likely it is that the data will yield a *random high* causing the trial being stopped.[25,66,68,69] 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 and stopping boundaries should always be *pre-specified* in the SAP.[70]

Trials may stop early for *futility*—the inability of a clinical trial to meet its objectives.[71] On the one hand stopping early for futility protects participants from exposure to ineffective treatment, saving resources for 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, contributing to the total evidence (in meta-analysis). Failure to report trials stopped early for futility leads to publication bias in future evidence syntheses.[72]

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 stopping boundary. Indeed, many trials are continued to conclusion despite clear evidence of harm. Statistical methods exist to assess 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.[71-73] The problem of stopping trials early for futility risks the opposite effect to stopping trials early for benefit. In stopping for futility, early results may represent *random lows*, which cause the illusion of no effect and the trial being stopped when, had more information been gathered, this *no effect signal* would disappear.

3.7 | Assessment of low frequency or long-term harms

RCTs rarely identify a pre-defined hypothesis to detect harms (as opposed to a hypothesis for efficacy), and are not powered to observe harms occurring infrequently or which only develop a considerable time after exposure.[74] Also, patients at highest risk of harm are often excluded from RCTs (e.g. older patients, those with multiple comorbidities, children), even if destined to become significant users of the treatment if licensed.[74] Additionally, in longer-term, larger RCTs it can be challenging to distinguish harm caused by treatment (iatrogenic) from that which is "inter-current and non-causal or just random error".[66] It is recognised, however, that rare harms may not become apparent until after a therapy has been licensed (Table 3 illustrates the number of patients to be observed to detect a given adverse event rate). In meta-analysis or systematic reviews, conclusions about harm

TABLE 3 Number of patients to be observed to detect a given adverse event, modified from[75]

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	1300
1:1000	3000	4800	6500
1:2000	6000	9600	13 000
1:10 000	30 000	48 000	65 000

may also be misleading if the available data are affected by publication bias.[74]

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

3.8 | A note about endpoints

RCTs report outcomes that may or may not be *clinically* meaningful. Particularly in early development, a drug's effect may be reported in terms of a surrogate, or proxy, for change in disease status (e.g. HbA1c — glycated haemoglobin — in diabetes). Changes to proxy markers are often described as a *soft* outcome. Whether a change in a proxy brought about by a drug translates into a meaningful effect on clinical (*hard*, unequivocal) outcomes can only be studied in larger, longer trials which allow for sufficient *hard* outcomes to accumulate in the treatment groups to detect a difference, if one exists.

While some surrogates appear to be directly correlated to *hard* outcomes, for example reducing systolic blood pressure has been well-established to reduce cardiovascular events (CVD).[76] The relationship between other proxies and *hard* outcomes is less clear, as for example, between HbA1c and complication outcomes in diabetes (e.g. CVD, amputation).

New drugs for the treatment of type 2 diabetes are licensed on the basis of reducing HbA1c and no signal of excess of cardiovascular or other safety events in meta-analysis of the available pre-licensing studies, with large-scale cardiovascular outcome trials (CVOT) *usually* undertaken post-licensing. Considering dipeptidyl peptidase-4 inhibitors and sodium-glucose transporter 2 inhibitors, it is possible to illustrate the problem with the relationship of proxies to clinical outcomes. Both recently licensed classes of drugs reduce HbA1c by improving glycaemic control. However, no dipeptidyl peptidase-4 inhibitor drugs appear to reduce the risk of CVD in CVOT (but do not increase the risk) despite reducing HbA1c. Conversely, all sodium-glucose transporter 2 inhibitor agents appear to reduce CVD risk to some extent in CVOT.[77] These disparate outcomes suggest that a reduction of HbA1c is not sufficient to predict whether a drug to treat diabetes will lead to a reduction in *hard* CVD clinical endpoints. Thus, careful consideration of whether *soft* endpoints reported in clinical trials translate into clinically meaningful *hard* endpoints must always be given.

4 | OBSERVATIONAL PHARMACOEPIDEMIOLOGY

Observational studies include cohort, case–control and cross-sectional studies.[78] Except in specific circumstances most OP studies should take the form of a cohort study.

The key difference between these types of study and RCTs is that, in OP studies, the intervention is selected for/by a patient, or the patient is selected by having been exposed to the intervention, rather than it being allocated randomly.[78] This makes it conceptually more difficult to attribute an outcome to a particular treatment and also introduces the potential that bias or confounding account for any differences observed.[78] In particular, an extremely challenging problem in OP is *allocation bias*. Also, the sensitivity and specificity of the outcome measures are often unknown in OP studies, so it is unclear if 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[79] in the UK) are an important part of post-licensing pharmacovigilance. With the increasing availability of electronic health 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 clinicians 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 large volumes of computerised observational data and a strong push to harness these.[3,4,80] 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).[81]

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

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

OP studies are often said to have high external validity.[4,12,82] 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 TEE detected in this subgroup of the OP study is in the same direction and order of magnitude as that reported in an RCT.[82] As such, OP studies can confirm and broaden the findings of RCTs to a wider population.[83]

STrengthening the Reporting of OBservational studies in Epidemiology (STROBE): like CONSORT, STROBE consists of guidelines and a

22-item checklist considered essential for good reporting.[84] Current guidelines date from 2009, which is the first iteration. They cover the 3 most commonly employed designs in observational studies: (i) cohort; (ii) case–control; and (iii) cross-sectional studies.[84] Also like CONSORT, adherence to STROBE may reduce bias and allow the reader to ascertain whether included analyses were preplanned or not, and if not why.[84] STROBE is more recent 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 (random sample of 100 studies in 2010), over 80% made appropriate use of STROBE.[85]

4.2 | 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 longer periods of time than RCTs. Indeed, some have been running for many decades, such as the Framingham Cardiovascular Cohort Study, operating for over 65 years.[86] This advantage of time means that observational studies are able to provide important data on patients' long-term experiences, particularly in the setting of chronic diseases with a natural history over many years.[87]

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 .05.[88] The extended period over which OP studies can be undertaken, and the relative ease of obtaining large enough population samples compared to RCTs, makes OP studies suited to the defining of adverse events and their incidence.[87] 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.[88] Observational studies can facilitate the detection of rare (<1/1000) and very rare (<1/10 000) adverse events (see Table 3) and are also able to provide long-term data on tolerability.[83,89]

Since observational studies frequently run in parallel with routine clinical care, they often cost less than RCTs.[90] In addition, OP studies might employ data available from clinical databases such as the Clinical Practice Research Datalink (UK), the Scottish Care Information—Diabetes Collaboration database and Health Maintenance Organization Research Network (USA).[91-93] Indeed, the future of OP is likely to be represented in such large longitudinal electronic healthcare record (disease registry or insurance provider) databases.

4.3 | OP studies can provide data to justify RCTs

OP studies often provide the evidence to justify, or to generate hypotheses for, an RCT[94] (see Table 4 for areas suited to observational studies). In addition, if the TEE detected in an OP study is very large then it is not always necessary to undertake an RCT.[94] There are multiple examples of treatments becoming established on the basis of observational data without confirmation in an RCT, such as, for example, the treatment of type 1 diabetes mellitus with insulin.[2]

TABLE 4 Particular areas suited to the use of observational studies[95]

Prospective evaluation of patient population and disease characteristics
Assessment and comparison of costs and effectiveness associated with diagnostics
Investigation of adherence to guidelines
Postmarketing surveillance
Detection of responsive subgroups
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

5 | 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.

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

The nonrandom allocation of patients in OP studies means that they are more prone to bias and confounding, both known and unknown.[96,97] Although strategies exist to mitigate the effects of these it is never possible to correct the results for all possible influences, particularly those unknown. Bradford Hill lists criteria for causal association[8] (Table 1), although, due to the inherent difficulty controlling for bias/confounding in OP studies, causality is more difficult to establish. Statistical association does not imply causality. However, the larger the TEE in OP studies the greater the support; yet stronger still if the observation of association is consistent in different studies/populations and with different study designs.[66,96-98]

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

The distortion caused by not randomising and blinding during an OP study has been associated with effect estimates as large or larger than the true treatment effect itself.[14] However, meta-analyses of the TEE in OP studies and RCTs have demonstrated that when good quality studies are analysed, the direction and magnitude is broadly

similar.[99-101] Nevertheless, the spectre of TEE over-inflation hangs over OP studies and should always be borne in mind when considering their results.

Many OP studies rely on data gathered through routine clinical practice. Conversely, for different reasons to those just described, this means that OP studies may also be at risk of under-estimation, where patients fail to seek healthcare and thus the true incident rate of a condition may not be recorded, or under-reporting, where following interaction with a healthcare system the data are inadequately reported. This under-estimation is a form of nondifferential information bias affecting the sensitivity and specificity of the outcome.[102]

5.3 | Approaches to deal with the limitations of OP

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

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

Incident-user design: this assumes that *both users and controls* have been identified by clinical staff as benefitting from a new prescription, making users and controls more similar, particularly in characteristics which may not be observable.[103] This does not always mean that incident-users and their controls are identical — for instance clinicians may avoid prescribing newly licensed drugs to frail 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*.[103]

Natural experiments: one example is *universal exposure* to avoid selection and allocation bias, where the exposure occurs in total populations rather than through choice, allowing comparisons to be made between exposed and unexposed time and causal inferences to be made.[104] 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. This allowed exploration of whether MMR 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.[105]

Another example of a natural experiment, devised as an alternative to RCTs, albeit applicable in limited circumstances, is *regression discontinuity design*. This uses a predetermined *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, who are then compared.[104] Assignment to intervention cannot be caused by the intervention but does require all participants to belong to the same population. The effect is measured by discontinuity from regression, which has been demonstrated mathematically to yield an unbiased estimate of a causal effect.[104]

Propensity scores: are designed to correct for the non-balanced distribution of characteristics between the exposed and unexposed groups and are more statistically efficient than multivariable regression models traditionally used to control for known confounders in OP studies.[106] However, like multivariable regression, propensity scores can only correct for known confounders rather than all confounders but, unlike multivariable regression the sensitivity of propensity scores for unknown confounders can be estimated and reported.[106] In order to develop an effective propensity score a thorough understanding of the covariates (i.e. the biology) is necessary for them to be included in score creation.[106] Propensity scores can be used for *matching* treated and untreated subjects, for *stratification* into mutually exclusive subsets, to create a synthetic sample in which the distribution of baseline covariates is independent of treatment assignment known as *inverse probability of treatment weighting* or for *covariate adjustment* where the outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score.[107]

Focussing on the dose–response relationship: one of Bradford Hill's criteria for causality is the presence of a dose–response relationship where one might expect, for example, to see a larger treatment effect from a larger exposure to an intervention. 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.[108] Using a 2-time updated exposure term, one for ever-/never-exposure and another for cumulative exposure, has been shown mathematically to remove the allocation bias from the cumulative exposure term and provide a more reliable TEE based on cumulative exposure.[109] This technique would yield a conservative TEE if exposure to an intervention caused an instantaneous, rather than gradual, change in risk.

Instrumental variable analysis: uses an *instrument* linked to the treatment, but not directly or indirectly linked to the outcome except via the treatment.[110] The challenge is instrument identification, which 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.[110,111] A good example to illustrate this might be differences in hospitals' formularies. In this case, the treatment's accessibility depends on inclusion in the hospital's formulary, satisfying the first assumption. Although patients are not randomly allocated to hospitals, it might be acceptable to assume that 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, e.g. quality-of-care, the instrument can be thought to affect outcome only via the treatment itself, satisfying the third assumption.[110,111]

This means that, in theory, by making these assumptions and collecting data on the instrument, it is possible to make TEEs on outcomes without having to adjust for confounders.[110,111]

Case cross-over design of case–control studies: the within-patient control design 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, assumptions must be met 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.[112-114]

Although in theory a case-cross over design could be used to investigate treatment effect, it is more often used to assess harm, such as demonstrating that recent vaccination does not appear to raise the risk of multiple sclerosis relapse.[115] Importantly, this type of study design cannot account for time-varying within-patient confounders, e.g. changes to body mass index. 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, nonprescribed doses would be omitted from the patient's prescribing record leading to information bias. This study design is also prone to recall bias, if patients' recollections are used to define exposure rather than more objective measures, such as prescriptions.[112-114]

Partial blinding: although most OP studies by their very nature do not utilise randomisation, it is still possible to employ some form of blinding. The published report should explain who was blinded and who was not as this helps with critical appraisal.[116]

5.5 | Missing data

Although missing data can occur in both RCTs and OP studies, RCTs often include protocols that go to great lengths to reduce this phenomenon. The collection of complete data may prove more challenging in OP studies, where data are collected through routine clinical

TABLE 5 Examples of various methods employed to handle missing data, modified from[117]

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 they are missing from elsewhere, existing values are used instead. This may lead to modelling problems where sample sizes and standard errors of covariates 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 no new 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 intrasubject variability and gives rise to an illusion of precision.
Maximum likelihood modelling	Assumes that the data present all arise from a multivariate 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 run on each iteration. Summary statistics are created by combining the statistic from each model run and is more robust as it retains the variability and uncertainty of the missing data.
Sensitivity analysis	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.

practice and often retrospectively. Missing data can lead to biased estimates particularly if they are not missing at random. There are a number of techniques detailed in Table 5 to handle missing data, although the best way to deal with data being missed is to prevent it from happening in the first place.[117]

6 | SOME SPECIFIC BIASES IN OP

The subsequent paragraphs give details about some specific biases in OP to consider.

Protopathic bias: an example would be pancreatic cancer causing diabetes, leading to the prescription of an antidiabetic 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 — a form of reverse causality.[127] 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.[128]

Surveillance (performance) bias in OP studies: an example might be the use of ultrasound Doppler 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 ultrasound Doppler in trauma patients.[129] 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.[24]

6.1 | 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 bias is remedied by ensuring that the *pre-exposure* time is counted, classified and analysed as *unexposed* person–time.[130-132]

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 on 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, but previously exposed to the first-line treatment), especially if there is a long lag between exposure and outcome. This can be avoided by comparing treatments in patients with similar disease duration/stage.[133]

Risk window bias: 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 events. It is best handled by sensitivity analysis comparing varying risk window durations.[134]

7 | CONCLUSIONS

OP studies and RCTs have both contributed substantially 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 RCTs reduce bias/confounding, meaning the effects detected are more likely to be caused by the treatment. However, RCTs do have shortcomings in relation to their generalisability 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* sub-group 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 sub-group 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 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.

COMPETING INTERESTS

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

The following authors have disclosed declarations of interest outside the submitted work: T.M.C. is funded by Diabetes UK with support from the British Heart Foundation. J.W.D. holds a grant from PledPharma AB. H.M.C. 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. H.M.C. is a shareholder in Bayer and Roche Pharmaceuticals. H.M.C. 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. D.J.W. is Vice Chair of MHRA but this paper is written in his capacity as a University of Edinburgh academic and the paper does not reflect the views of MHRA.

Glossary (A-Z)	
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.[96,97] These systematic differences between how participants *are assigned to their treatment group*, means 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.
Allocation concealment	Hiding the sequence of allocation prior to recruitment, so it is not possible to predict to which treatment group a participant will be assigned.[118]
Attrition bias	The unequal loss of participants between the treatment groups such that they are no longer similar to one another.[119,120] It is a type of after-the-event selection bias, where one group (or both) are no longer representative of the condition under study
Basket trials	A type of adaptive RCT. 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 subtrial (drug–mutation pair testing) with matched therapy or control.
Bias	"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".[121,122] In general, you cannot adjust for bias in an analysis. Bias leads to spurious (untrue) associations.
Blinding (masking)	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.[123] The effect of blinding is to reduce observer bias in ascertaining the outcomes of interest, a form of differential information bias.
Case–control studies	Retrospective, where 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 afterwards.[78]
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.
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 outcome over time.[78]
Cluster RCTs (cRCTs)	Randomise at the group-level, say a clinic or hospital, rather than at the individual patient-level. 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 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.
Confounding	Occurs when an apparent association between an exposure of interest and an outcome is due to another factor that is associated with both the exposure and also independently with the outcome but is not in the causal pathway between the two.[16] Confounding differs from bias in that, if the confounder is known, statistical methods can often 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.[124] It is of course, not possible to correct for unknown confounders. Confounding leads to true, but indirect (not causal), associations.
CONSORT (CONsolidated Standards Of Reporting Trials)[18]	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.
Crossover RCTs (xRCTs):	A within-subject study design, where participants are randomly exposed to interventions in sequence (treatment A followed by treatment B or *vice versa*), and thus act as their own controls.[50] One of the treatments may be placebo or an active control. xRCTs can give greater precision of treatment effect, given the same number of subjects, than a similarly sized parallel group study.
Cross-sectional studies	Look at the prevalence of a disease at a specific time point and may use historical data to establish exposure.[78]
Effect modifier	Is a clinical characteristic (e.g. age, sex, genotype) that 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 >30 kg/m^2, thus in this context body mass index can be considered an effect modifier).[7]
Efficacy	"The performance of an intervention under ideal, controlled circumstances *compared to placebo*."[4,6]

(Continued)

Glossary (A-Z)	
(Relative) effectiveness	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.[4,6]
Efficacy-effectiveness gap	The inconsistency between the effects of an intervention reported in clinical trials compared to that reported in routine clinical practice.[4]
External validity	The extent to which the findings of a study are valid outside the context of the study. A study with good external validity is likely to have results which apply to a broad range of people with heterogeneous characteristics, which are largely generalisable (similar to generalisability).[4,12,82]
Factorial RCT (fRCT):	Allow for the assessment of multiple treatments in the same population, maximises study power and also provides information on interactions between treatments.[53] In its simplest form, a *2×2* 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.[53,125]
Generalisability	Whether study results apply to the population in whom they will be applied. It relates to the degree to which a treatment effect estimate can be applied to a wide group of patients under *usual conditions* (it is a similar concept to external validity).
Immortal time bias	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.
Incident-user design	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.
Information bias	Occurs when information is obtained differently between exposed and unexposed cases such as a flaw in measuring exposure, outcomes or covariates with differing accuracy between groups. For continuous variables this is known as measurement error, for discrete variables classification error.[96,97] 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. In nondifferential 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.
Intention to treat analysis (ITT)	When participants are analysed in the group to which they were assigned, irrespective of whether they completed the study.
Internal validity	The extent to which causal conclusions regarding a study are justified.[25] 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]
Large simple RCTs (lsRCTs):	pRCTs (see below) but with protocols mandating only minimal data collection on outcomes important to patients or care providers.
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.
Multi-arm RCTs	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 2-arm trials investigating the same question and provide data for direct comparison rather than many 2-arm studies being compared in meta-analysis, which causes difficulties in interpretation when the studies are heterogeneous.[56,57] It may also be the case that multi-arm trials recruit more effectively than 2-arm trials, possibly since the multiple arms, with different inclusion criteria, mean more patients are eligible, and well-designed multi-arm studies may provide significant patient benefit compared to multiple 2-armed trials.[56]
Natural experiments	Alternatives to RCTs that utilise naturally occurring circumstances to separate variables that usually associate together in a before and after cohort study.[64]
Open-label RCT	An RCT where allocation concealment is undertaken but the study is not blinded and may 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.[118]

(Continued)

Glossary (A-Z)	
Partial blinding	Involves the blinding of some aspects of an OP study (or indeed an RCT [see open-label RCT]), for example observer blinding. The preferred technique is to separate the extraction of exposure information from outcome information.[116]
Platform trials	A type of adaptive RCT. They 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.[58]
Pragmatic RCTs: (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.[37]
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 nonoverlapping data from analysis on the basis of an understanding of covariates that affect the condition being studied.[106]
Prospective randomised open blinded endpoint study (PROBE)	A particular type of open-label study design thought to be more cost-effective than the double-blind prospective study. It uses strict randomisation and hard endpoint definitions (ones that are well-defined and measured objectively) to allow for the comparison of interventions to take place.
Protopathic bias	Occurs when the prescription of a treatment is caused by the symptoms of an undiagnosed condition.
Publication bias	When publication depends on the hypothesis being tested and the significance and direction of the effects detected.[119,120] A type of differential information bias.
Randomisation	The random allocation of participants to intervention groups and achieves comparability between these, especially in terms of prior risk of the outcome of interest and any effect modifiers. Randomisation allows causal inferences to be made; in other words, the treatment effect observed is probably due to the intervention, all things being equal.
Randomised control trial (RCT)	A study in which a number of similar people are randomly assigned to 2 (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.
Randomised database studies	A specific form of IsRCT, 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 such as participant consent are still to be standardised.
Relative efficacy	Similar to *efficacy* except *comparison is to a standard alternative* rather than placebo.[4,6]
Risk window bias	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.
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, particularly in terms of prior risk of the outcome of interest or effect modifiers.[126] 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*.
Statistical analysis plan (SAP)	A "more detailed and technical elaboration of the principal features of analysis included in the trial protocol"[22]
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".[84]
Stratified randomisation	If certain covariates 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 that there be equal numbers of patients with a rare, severe disease phenotype in both arms.[10]
Surveillance bias (detection bias)	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.
Umbrella trials	A type of adaptive RCT. A single class or type of tumour is molecularly screened and assigned to subtrials 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).

ACKNOWLEDGEMENTS

TMC is a Diabetes UK 'Sir George Alberti Clinical Research Fellow' (Grant number: 18/0005786), although the views represented in this article are his own and not those of Diabetes UK.

ORCID

Thomas M. Caparrotta https://orcid.org/0000-0001-9009-9179
James W. Dear https://orcid.org/0000-0002-8630-8625
Helen M. Colhoun https://orcid.org/0000-0002-8345-3288
David J. Webb https://orcid.org/0000-0003-0755-1756

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. Sources of evidence for assessing the safety, efficacy and effectiveness of medicines. The Academy of Medical Sciences. Available at: https://acmedsci.ac.uk/policy/policy-projects/methods-of-evaluating-evidence. Accessed January 8, 2019.
4. IMI GetReal - Real-Life Data in Drug Development > Home. Available at: http://www.imi-getreal.eu/. Accessed January 8, 2019.
5. Akobeng AK. Understanding randomised controlled trials. *Arch Dis Child*. 2005;90(8):840-844.
6. Singal AG, Higgins PDR, Waljee AK. A primer on effectiveness and efficacy trials. *Clin Transl Gastroenterol*. 2014;5(1):e45.
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.
8. Hill AB. The environment and disease: association or causation? *Proc R Soc Med*. 1965;58:295-300.
9. Torgerson DJ, Roberts C. Randomisation methods: concealment. *BMJ*. 1999;319(7206):375-376.
10. Gordis L. Assessing preventive and therapeutic measures: randomized trials, assessing preventive and therapeutic measures: randomized trials. In: *Epidemiology*. 5th ed. Philadelphia, Pennsylvania: Elsevier/Saunders; 2014:138-154.
11. Shrier I, Boivin JF, 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.
12. Cartwright N. What are randomised controlled trials good for? *Philos Stud*. 2010;147(1):59-70.
13. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. *Fam Med*. 2007;39(2):132-137.
14. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317(7167):1185-1190.
15. 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.
16. Sedgwick P. Randomised controlled trials: understanding confounding. *BMJ*. 2015;351:h5119.
17. Cochrane Handbook for Systematic Reviews of Interventions. Available at: http://handbook.cochrane.org/. Accessed May 13, 2017.
18. Schulz KF. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*. 2010;152(11):726-732.
19. 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. *Cochrane Database Syst Rev*. 2012;11:MR000030.
20. 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.
21. Plint AC, Moher D, Morrison A, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust Pyrmont*. 2006;185:263-267.
22. Statistical Principles for Clinical Trials: ICH. Available at: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html. Accessed January 8, 2019.
23. 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.
24. ENCePP Home Page. Available at: http://www.encepp.eu/standards_and_guidances/methodologicalGuide4.shtml. Accessed January 8, 2019.
25. Behi R, Nolan M. Causality and control: threats to internal validity. *Br J Nurs*. 1996;5(6):374-377.
26. Levin KA. Study design VII. Randomised controlled trials. *Evid Based Dent*. 2007;8(1):22-23.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. Woodcock J, Ware JH, Miller PW, McMurray JJV, Harrington DP, Drazen JM. Clinical trials series. *N Engl J Med*. 2016;374:2167-2167.
35. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*. 2005;365(9453):82-93.
36. Dumville JC, Torgerson DJ, Hewitt CE. Research methods: reporting attrition in randomised controlled trials. *BMJ*. 2006;332(7547):969-971.

37. Hotopf M. The pragmatic randomised controlled trial. *Adv Psychiatr Treat*. 2002;8(5):326-333.
38. 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.
39. 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.
40. Fröbert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. *NEJM*. 2013;369(17):1587-1597.
41. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. *Stat Med*. 2002;21(11):1503-1511.
42. Abraham, J. International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use. In *Handbook of Transnational Economic Governance Regimes* (eds. Brouder, A. & Tietje, C.) 1041–1054. Leiden, Netherlands: Brill, 2009. https://doi.org/10.1163/ej.9789004163300.i-1081.897
43. MRC/Wellcome Trust Workshop: regulation and biomedical research. (2008).
44. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*. 1996;312(7040):1215-1218.
45. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. *Am J Public Health*. 2004;94(3):416-422.
46. 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.
47. Joffe S. Evaluating novel therapies during the Ebola epidemic. *JAMA*. 2014;312(13):1299-1300.
48. 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.
49. 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.
50. Sibbald B, Roberts C. Understanding controlled trials crossover trials. *BMJ*. 1998;316(7146):1719-1720.
51. Sedgwick P. What is a crossover trial? *BMJ*. 2014;348:g3191.
52. Dwosh IL, Giles AR, Ford PM, Pater JL, Anastassiades TP. Plasmapheresis therapy in rheumatoid arthritis. *N Engl J Med*. 1983;308(19): 1124-1129.
53. Torgerson DJ, Torgerson C. Factorial RCTs. In: *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*. (pp. 114-118). Basingstoke, Hampshire: Palgrave Macmillan Limited; 2008.
54. 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.
55. 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.
56. 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: 849-851.
57. Parmar MKB, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. *Lancet*. 2014;384(9940):283-284.
58. 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.
59. The STAMPEDE Trial. *STAMPEDE* Available at: http://www.stampedetrial.org/. Accessed January 8, 2019.
60. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Stat Med*. 2014;33(19):3269-3279.
61. Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *NEJM*. 2017;377(7):644-657.
62. Kasenda B, Schandelmaier S, Sun X, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. *BMJ*. 2014;349(1):g4539.
63. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344(1):e1553-e1553.
64. Rutter M, Pickles A. Annual research review: threats to the validity of child psychiatry and psychology. *J Child Psychol Psychiatry*. 2016;57(3):398-416.
65. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*. 2011;22(11):1359-1366.
66. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet*. 2008;372(9656):2152-2161.
67. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991;266(1):93-98.
68. Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin Trials*. 2009;6(2):119-125.
69. Guyatt GH, Briel M, Glasziou P, Bassler D, Montori VM. Problems of stopping trials early. *BMJ*. 2012;344(1):e3863.
70. Skovlund E. Repeated significance tests on accumulating survival data. *J Clin Epidemiol*. 1999;52(11):1083-1088.
71. Snapinn S, Chen M-G, Jiang Q, Koutsoukos T. Assessment of futility in clinical trials. *Pharm Stat*. 2006;5(4):273-281.
72. Pocock SJ. When to stop a clinical trial. *BMJ*. 1992;305(6847): 235-240.
73. Pocock SJ. Current controversies in data monitoring for clinical trials. *Clin Trials J Soc Clin Trials*. 2006;3(6):513-521.
74. Chou R, Aronson N, Atkins D, et al. Assessing harms when comparing medical interventions. In: *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville, Maryland: Agency for Healthcare Research and Quality (US); 2008.
75. Heneghan C. Rare adverse events in clinical trials: understanding the rule of three. *BMJ EBM Spotlight* (2017). Available at: http://blogs.bmj.com/bmjebmspotlight/2017/11/14/rare-adverse-events-clinical-trials-understanding-rule-three/. Accessed January 8, 2019.
76. Kannel WB. Elevated systolic blood pressure as a cardiovascular risk factor. *Am J Cardiol*. 2000;85(2):251-255.
77. Home P. Cardiovascular outcome trials of glucose-lowering medications: an update. *Diabetologia*. 2019;62(3):357-369.
78. Rosenbaum, P. R. Observational Studies. In *Observational Studies* 1–17 (Springer New York, 2002). https://doi.org/10.1007/978-1-4757-3692-2_1

79. Yellow Card Scheme - MHRA. Available at: https://yellowcard.mhra.gov.uk/. Accessed April 23, 2019.
80. Medical Research Council, M. R. C. Health Data Research UK (HDR UK) (2017). Available at: https://www.mrc.ac.uk/about/institutes-units-centres/uk-institute-for-health-and-biomedical-informatics-research/. Accessed January 8, 2019.
81. 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(1):c693.
82. Steckler A, McLeroy KR. The importance of external validity. *Am J Public Health*. 2008;98(1):9-10.
83. Silverman SL. From randomized controlled trials to observational studies. *Am J Med*. 2009;122(2):114-120.
84. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-349.
85. da Costa BR, Cevallos M, Altman DG, Rutjes AWS, Egger M. Uses and misuses of the STROBE statement: bibliographic study. *BMJ Open*. 2011;1:e000048.
86. 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.
87. 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.
88. 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.
89. Chan EW, Liu KQL, Chui CSL, Sing CW, Wong LYL, Wong ICK. Adverse drug reactions—examples of detection of rare events using databases. *Br J Clin Pharmacol*. 2015;80(4):855-861.
90. 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.
91. Clinical Practice Research Datalink - CPRD. Available at: https://www.cprd.com/intro.asp. Accessed January 8, 2019.
92. SCI-Diabetes. Available at: http://www.sci-diabetes.scot.nhs.uk/. Accessed January 8, 2019.
93. Health Maintenance Organization Research Network (HMORN) UCSF Center for Diabetes Translational Research|Global Research Projects. Available at: https://globalprojects.ucsf.edu/project/health-maintenance-organization-research-network-hmorn-ucsf-center-diabetes-translational. Accessed January 8, 2019.
94. 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(1):51.
95. Tavazzi L. Do we need clinical registries? *Eur Heart J*. 2014;35(1):7-9.
96. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359(9302):248-252.
97. Ranstam J. Bias in observational studies. *Acta Radiol*. 2008;49(6):644-645.
98. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? *NEJM*. 2000;342(25):1907-1909.
99. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *NEJM*. 2000;342(25):1887-1892.
100. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *NEJM*. 2000;342(25):1878-1886.
101. Anglemyer, A., Horvath, H. T. & Bero, L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. In *Cochrane Database of Systematic Reviews* (ed. The Cochrane Collaboration) (John Wiley & Sons, Ltd, 2014). https://doi.org/10.1002/14651858.MR000034.pub2
102. Gibbons CL, Mangen MJJ, Plass D, et al. Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*. 2014;14(147).
103. Johnson ES, Bartman BA, Briesacher BA, et al. The incident user design in comparative effectiveness research. *Pharmacoepidemiol Drug Saf*. 2013;22(1):1-6.
104. Rutter M. Proceeding from observed correlation to causal inference: the use of natural experiments. *Perspect Psychol Sci*. 2007;2(4):377-395.
105. 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.
106. Okoli GN, Sanders RD, Myles P. Demystifying propensity scores. *Br J Anaesth*. 2014;112(1):13-15.
107. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46(3):399-424.
108. 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.
109. 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.
110. Hernán MA, Robins JM. Instruments for causal inference: an Epidemiologist's dream? *Epidemiology*. 2006;17(4):360-372.
111. Klungel OH, Uddin MJ, de Boer A, Belitser SV, Groenwold RH, Roes KC. Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*. 2015;06:2.
112. Etminan M, Samii A. Pharmacoepidemiology I: a review of Pharmacoepidemiologic study designs. *Pharmacotherapy*. 2004;24(8):964-969.
113. 'Chris' Delaney JA, Suissa S. The case-crossover study design in pharmacoepidemiology. *Stat Methods Med Res*. 2009;18:53-65.
114. Donnan PT, Wang J. The case-crossover and case-time-control designs in pharmacoepidemiology. *Pharmacoepidemiol Drug Saf*. 2001;10(3):259-262.
115. 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.
116. Parker, R. A. & Berman, N. G. Chapter 28 - Blinding in observational studies. in *Planning Clinical Research* (Cambridge University Press, 2016). https://doi.org/10.1017/CBO9781139024716
117. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406.
118. Sedgwick P. What is an open label trial? *BMJ*. 2014;348:g3434.
119. Viswanathan M, Berkman ND, Dryden DM, Hartling L. *Approaches to Assessing the Risk of Bias in Studies*. Rockville, Maryland: Agency for Healthcare Research and Quality (US); 2013.
120. Hammer GP, du Prel J-B, Blettner M. Avoiding bias in observational studies: part 8 in a series of articles on evaluation of scientific publications. *Dtsch Ärztebl Int*. 2009;106:664.
121. A Word About Evidence: 6. Bias—a proposed definition. *Catalog of Bias* (2018). Available at: https://catalogofbias.org/2018/06/15/a-

word-about-evidence-6-bias-a-proposed-definition/. Accessed January 8, 2019.

122. Porta M. *Dictionary of Epidemiology*. New York, New York: Oxford University Press, Incorporated; 2014.
123. Day SJ, Altman DG. Blinding in clinical trials and other studies. *BMJ*. 2000;321(7259):504.
124. Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ*. 2000;321(7256):255-256.
125. Sedgwick P. What is a factorial study design? *BMJ*. 2014;349:g5455.
126. Sedgwick P. Selection bias versus allocation bias. *BMJ*. 2013;346:f3345.
127. Korhonen MJ, Huupponen R, Ruokoniemi P, Helin-Salmivaara A. Protopathic bias in observational studies on statin effectiveness. *Eur J Clin Pharmacol*. 2009;65(11):1167-1168.
128. Shin JY. *Potential overestimation of risk by protopathic bias and mitigation by the introduction of lag-time*. 2016;354:i4857.
129. Haut ER, Pronovost PJ. Surveillance bias in outcomes reporting. *JAMA*. 2011;305(23):2462-2463.
130. Agarwal P, Moshier E, Ru M, et al. Immortal time bias in observational studies of time-to-event outcomes. *Cancer Control J Moffitt Cancer Cent*. 2018;36:195-199.
131. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ*. 2010;340(907–911):b5087.
132. Suissa S. Immortal time bias in Pharmacoepidemiology. *Am J Epidemiol*. 2008;167(4):492-499.
133. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care*. 2012;35(12):2665-2673.
134. van Staa TP, Abenhaim L, Leufkens H. A study of the effects of exposure misclassification due to the time-window design in pharmacoepidemiologic studies. *J Clin Epidemiol*. 1994;47(2):183-189.

How to cite this article: Caparrotta TM, Dear JM, Colhoun HM, Webb DJ. Pharmacoepidemiology: Using randomised control trials and observational studies in clinical decision-making. *Br J Clin Pharmacol*. 2019;85:1907–1924. https://doi.org/10.1111/bcp.14024