OPEN ACCESS

Evidence synthesis

Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making

Grammati Sarri ,[1] Elisabetta Patorno ,[2] Hongbo Yuan,[3] Jianfei (Jeff) Guo,[4] Dimitri Bennett ,[5] Xuerong Wen,[6] Andrew R Zullo ,[7] Joan Largent,[8] Mary Panaccio,[9] Mugdha Gokhale,[10] Daniela Claudia Moga,[11] M Sanni Ali,[12,13,14] Thomas P A Debray [15,16]

10.1136/bmjebm-2020-111493

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjebm-2020-111493).

For numbered affiliations see end of article.

Correspondence to:
Dr Grammati Sarri, Visible Analytics, Oxford OX2 0DP, UK; grammati.sarri@visibleanalytics.co.uk

Check for updates

To cite: Sarri G, Patorno E, Yuan H, *et al*. *BMJ Evidence-Based Medicine* 2022;**27**:109–119.

Abstract

Introduction: High-quality randomised controlled trials (RCTs) provide the most reliable evidence on the comparative efficacy of new medicines. However, non-randomised studies (NRS) are increasingly recognised as a source of insights into the real-world performance of novel therapeutic products, particularly when traditional RCTs are impractical or lack generalisability. This means there is a growing need for synthesising evidence from RCTs and NRS in healthcare decision making, particularly given recent developments such as innovative study designs, digital technologies and linked databases across countries. Crucially, however, no formal framework exists to guide the integration of these data types. Objectives and Methods: To address this gap, we used a mixed methods approach (review of existing guidance, methodological papers, Delphi survey) to develop guidance for researchers and healthcare decision-makers on when and how to best combine evidence from NRS and RCTs to improve transparency and build confidence in the resulting summary effect estimates. Results: Our framework comprises seven steps on guiding the integration and interpretation of evidence from NRS and RCTs and we offer recommendations on the most appropriate statistical approaches based on three main analytical scenarios in healthcare decision making (specifically, 'high-bar evidence' when RCTs are the preferred source of evidence, 'medium,' and 'low' when NRS is the main source of inference). Conclusion: Our framework augments existing guidance on assessing the quality of NRS and their compatibility with RCTs for evidence synthesis, while also highlighting potential challenges in implementing it. This manuscript received endorsement from the International Society for Pharmacoepidemiology.

Introduction

Comparative effectiveness research is a key step in the evaluation of novel therapeutic products. Although randomised controlled clinical trials (RCTs) are the established method for providing information on the relative efficacy and safety of health interventions, it may be impractical to conduct them, and those available may be sparse, small and potentially unrepresentative of the patient populations or conditions found in real-world settings. Consequently, evidence from such studies alone might not reliably reflect how medical interventions are likely to perform when used in everyday clinical care.[1–3] For this reason, there has been a growing demand, especially from regulatory bodies (Food and Drug Administration [FDA], European Medicines Agency [EMA]) to incorporate real-world evidence (RWE) from routine clinical practice as found in non-randomised studies (NRS) to complement information from RCTs and potentially cover the 'efficacy-effectiveness' gap.[4–7] The regulatory acceptance of RWE will present the challenge to other healthcare decision-makers (payers, health technology assessment (HTA) bodies) to increasingly use NRS for their policy decisions. Such evidence is potentially available via healthcare claims databases, electronic health records (EHR), patient registries,[8–10] and cohort and case–control studies, facilitated by the emergence of digital technologies,[9] and the promotion of exchange of EHRs across countries.[9] These changes have occurred in parallel with increasing pressure from patient advocacy groups to consider more patient-centred information in health products value assessments.[11]

Need for guidance

The International Society for Pharmacoepidemiology (ISPE) Comparative Effectiveness Research (CER) Special Interest Group (SIG) has previously commented on the challenges of using RWE from NRS in assessing comparative treatment effects. It has also highlighted how recent methodological advances can help to address inherent limitations of NRS, such as selection and confounding.[12] Recent publications have emphasised the need for ongoing discussion among stakeholders about when and how data from NRS can be used when the 'totality of evidence' is considered for assessing

Summary box

What is already known about this subject?

- Non-randomised studies (NRS) are increasingly recognised as being complementary to randomised controlled trial evidence for making credible estimates of the comparative treatment effects of medical products.
- The lack of methodological frameworks to guide synthesis of results of NRS with those of randomised clinical trials (RCTs) is a major cause of the low uptake of cross-study design synthesis for healthcare decision making and has been widely recognised by different organisations.
- "What can our framework offer"?
- We propose a seven-step framework to systematically identify evidence for NRS, critically appraise and appropriately synthesise it with the results from RCTs.
- Our framework considers three main analytical scenarios based on the evidence-generation needs for a healthcare decision-making problem; 'high-bar,' 'medium' and 'low' depending on whether evidence from randomised trials or non-randomised studies is the main source for trustworthy summary treatment effect estimates.
- Our framework emphasizes that the effect estimates from all the randomised and non-randomised evidence should not directly be combined in a meta-analysis without any type of statistical adjustment. When cross-design synthesis is considered appropriate, our framework guides researchers to select the most relevant statistical technique for an analytical scenario, such as using evidence from non-randomised studies as priors, in three-level hierarchical models and in bias-adjusted analysis. Expert clinical opinion and statistical expertise is required to avoid misleading results from combined analysis of non-randomised and randomised studies and increasing the risk of poorly informed healthcare decisions with harmful consequences to patients.
- "How might this framework impact healthcare decision-making in the future?"
- This framework will ultimately facilitate decisions around if, when and how evidence from NRS can be combined along RCTs and produce reliable treatment estimates applicable to a specific targeted population relevant for healthcare decision-making.

medical products, including complementing RCTs, to strengthen evidence packages for novel treatments.[13–15] However, there is a lack of methodological guidance on selection, appraisal and synthesis of evidence across different study designs in a consistent and reproducible manner. Other researchers are working on similar frameworks with a focus on specific conditions, such as cancer.[16] This methodological gap has been a key cause of the scepticism of regulators and healthcare decision-makers towards adopting novel methodologies proposed for the analysis of NRS.[7] Our proposed, comprehensive framework provides much-needed guidance to fill in these knowledge gaps to ensure the validity of non-RCT results.

Summary points of the framework

This framework is intended for use when NRS is considered for CER (or otherwise called relative effectiveness assessment in the EU HTA context) to address limitations of RCTs for licensing applications for primary conditional or secondary approvals in other indications, or to provide additional information for regulatory or reimbursement decisions for existing (standard of care) treatments.[17] For instance, the framework could be relevant for rare diseases, in which conducting traditional RCTs may be impractical (eg, due to recruitment difficulties). It might also facilitate assessment of real-world performance of products in patients with multiple comorbidities or at longer time points.

For the purposes of the framework, NRS is defined as those where the assignment of patients to a therapeutic product is not determined by a trial protocol; where additional diagnostic or monitoring procedures are not used or do not influence the care patients receive but instead represent routine clinical practice.[3 10] It is also assumed that NRS data can be collected either prospectively or retrospectively by observation in routine clinical practice, and can be analysed using epidemiological (biostatistical) methods.

Framework aims and development

The goal for our framework (figure 1) is to enable the trustworthy generation of results from combining NRS and RCT data, by providing specific recommendations on the appraisal tools of study quality, how to select the most reliable NRS evidence for a quantitative analysis with RCTs and various statistical approaches. More specifically, it comprises seven steps, some of which are well-established processes in evidence-based medicine (eg, systematic search and identification of relevant evidence (steps 1 and 2)) and, as such, are not described in full herein (readers should follow the guidance by Cochrane,[18] the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[19] and the Strengthening the Reporting of Observational Studies in Epidemiology Statement[20]). The goal is to provide specific recommendations for the critical appraisal of NRS (steps 3 and 4), for the implementation of statistical approaches to combine the results from NRS and RCTs (steps 5 and 6), and for facilitating a reliable interpretation of pooled (meta-analysed) results (step 7). For that reason, a mixed-methods approach was adopted for retrieving the most relevant literature and capitalising on the multidisciplinary experience of the working group on pharmacoepidemiology, observational statistical analysis and healthcare decision making. For step 3, we conducted a systematic literature review following PRISMA guidelines and searching indexed databases (Embase, PubMed) and general websites for tools that evaluated the validity of NRS from inception to November 2019 (online supplemental table 1) and online supplemental figures 1 and 23). In addition, a Delphi survey among the ISPE CER SIG was conducted to identify the main critical elements that can threaten the validity of NRS and developed the evaluation framework for assessing the validity of existing tools (Supplementary figure 2). For steps 4–7, we used a snowballing approach to perform reference checking of relevant publications (already known to the working group) from previous or ongoing RWE initiatives and key organisations (such as the Innovative Medicines Initiative [IMI] GetReal, FDA RWE Framework, Institute for Clinical and Economic Review [ICER], EMA, Duke-Margolis Institute, HTA bodies, International Society for Pharmacoeconomics and Outcomes Research ([ISPOR]/

Figure 1 International Society for Pharmacoepidemiology (ISPE) CER SIG framework for combining NRS with RCTs. CER, comparative effectiveness research; NRS, non-randomised studies; RCT, randomised controlled trial; SIG, Special Interest Group.

ISPE endorsed publications) and publications from selected journals (such as Clinical Pharmacology and Therapeutics, Research Synthesis Method and Statistics in Medicine).

Steps of a systematic review combining NRS and RCTs (1–4)

Step 1: defining the research question and identifying systematically the evidence

The identification and synthesis of all available relevant evidence (RCTs or NRS) in healthcare decision making must be done in a systematic, reproducible and rigorous way to ensure unbiased results regarding the effectiveness and safety of medical products. Accordingly, we recommend that, before any quantitative analysis comparing effects of different medical treatments is considered, researchers and healthcare decision-makers should specify a clear research question that defines the scope ('conceptual step'[21]), the Population, Intervention, Comparison, Outcomes, Study design and Time (PICOST) criteria. The PICOST can be used to conduct a scoping literature review and determine the need for summarising evidence across RCTs and NRS. It is, for instance, possible that published RCTs are scarce, or do not provide much information on important outcomes (eg, when serious harms of a medical treatment are rare or do not occur during the RCT follow-up). This decision may depend on both the frequency of an outcome but also on its importance/weight for the decision making.[18]

When setting the PICOST criteria, it is advisable to search the COMET database and record if a core outcome set is available for the condition of interest. Additional searches, such as the Outcome Measures Framework by the Agency for Healthcare Research and Quality and recent movements by EU IMI2 initiatives (Big Data for Better Outcomes) may facilitate the selection of standardised, measurable real-world outcomes.

When defining the research question, reviewers should also prespecify a list of 'core' confounders for which adjustment is deemed necessary in NRS. 'Core' confounders are defined as measured variables that influence treatment assignment, are predictive of the outcome and remove confounding when adjusted for. It is also helpful to identify, at this stage, (eg, intermediate or collider) variables that should not be adjusted for in NRS. A practical approach for preselecting 'core' confounders is to leverage prior knowledge of causal relationships for the specific decision problem (eg, by constructing causal diagrams[22]) and/or eliciting expert clinical opinion.

We advise readers to follow the detailed guidance by the Cochrane Collaboration (chapter 24) on this topic and apply additional search strategies to overcome specific challenges associated with the identification of NRS (eg, insufficient indexing of older NRS, large volume of evidence retrieved, additional time and resources for searching, identification of multiple publications and avoidance of 'duplicate' data set analyses).[23–25]

Step 2: data extraction

This critical step of the framework will largely determine the availability of key information, and therefore, the selection of NRS to be considered in the quantitative synthesis of evidence across study designs (step 5 in the framework).[26–28] Well-established data collection processes such as using a predefined data extraction template and dual extraction by two independent reviewers should be followed.[25] Incomplete data have been widely recognised as an important challenge when NRS are used in CER. The ability to link databases is a useful way to fill any data gaps but also to validate the data, therefore related datasets should be carefully cross referenced and extracted.[29–31] In general, reporting of information for each NRS should follow the same principles as the extraction of RCTs; information on study design, population, interventions, types of analyses and summary treatment effect statistics (such as extracting of treatment effect estimates using time-to event models and avoiding binary outcomes)[30–32] Additional data should be extracted to facilitate the assessment of different type of biases (eg, selection, attrition bias, outcome reporting bias). For instance, it is recommended to extract a list of confounders considered for the adjusted treatment effect analyses, or the method of propensity score adjustment.

With regard to extraction of summary effect estimates, when adopting non-collapsible effect measures such as ORs or HRs, it is important to distinguish between marginal (ie, population average unadjusted) and conditional (ie, covariate-adjusted) treatment effects.[33 34] Marginal effect measures greatly depend on the distribution of patient characteristics, and may vary even in the absence of confounding.[35] Previous research has shown that the difference between marginal and conditional effects can be substantial, especially when the number of prognostic factors exceeds five, the OR is above 1.25 (or smaller than 0.8), or the incidence proportion is between 0.05 and 0.95.[34] For this reason, pooling of marginal OR or HR estimates in such situations is not recommended. Further, when the marginal effect sizes are of primary interest, it may be

Box 1 Methodological challenges to be addressed by quality tools for non-randomised studies

- Methods for selecting participants (sampling strategies to correct selection bias, inclusion and exclusion criteria of target population, depletion of susceptibles, external validity of target population).
- Definition and measurement of exposure, outcomes, covariates and follow-up.
- Methods to address specific sources of bias through study design (new user design, active comparator design, methods to correct for immortal time bias or time-window bias, detection or surveillance bias, lost to follow-up bias, non-contemporaneous comparator bias, reverse causation, misclassification bias).
- Confounding (study design to *minimise* confounding, key confounders measured and included in statistical analysis, potential unmeasured confounding addressed in the analysis (please see online supplemental figure 4) for a summary of methods to adjust for either known or unknown confounding).
- Lack of appropriateness of statistical analyses (with specific mention of overadjustment, and/or incorrect outcome model specification).
- Methods for assessing statistical uncertainty in the findings.
- Methods for assessing internal validity (eg, sensitivity analysis addressing potential confounding, measurement error or other biases).
- Methods for assessing external validity (eg, post hoc subgroup analysis, validation of results with other similar population).

helpful to distinguish between the average treatment effect in the entire targeted population and the average treatment effect on the treated group in the study. These estimands target different populations or subgroups within the same population, and therefore can yield different treatment estimates.[36] The relevance of (differences in) estimands is further discussed in steps 5–7. Finally, estimates that are not directly available from the publication can sometimes be derived from other reported information.[30–32]

Step 3: critical appraisal of available data sources

Following previous guidance,[37 38] the group strongly recommends that both RCTs and NRS should rigorously be assessed for their validity and credibility before any cross-design synthesis can be considered. Results from the Delphi survey which was conducted as part of this framework development identified the following methodological challenges most associated with NRS (box 1).

Although tools for critical appraisal are widely available,[39 40] they vary considerably in their content (quality topics covered). The choice of appraisal tool is therefore a concern, as it may affect the selection of NRS for quantitative analysis and credibility of subsequent meta-analysis results. We recently conducted a systematic review to evaluate existing tools for critical appraisal of NRS and found that most of these cover the critical quality domains (box 1). Unfortunately, items to identify some fatal methodological flaws (eg, inability to conduct a study using new-user design or active comparator design, immortal time bias, depletion of susceptibles, reverse causation), and assessing issues around the internal and external validity of NRS results are currently missing in most of the existing tools. Based on our findings, we recommend ROBINS-I and GRACE as these tools cover most issues that are commonly encountered in NRS. However, it is advised to perform a supplementary assessment on the domains not fully covered by these tools (online supplemental figure 3). Tools for assessing RCTs have been reviewed previously and the use of Cochrane Risk of Bias tool is recommended,[41] as use of this is already an established practice in assessing the quality of RCTs.

Step 4: other issues to consider: small-study effects and publication bias

Critical appraisal tools may help to discover important limitations of NRS and RCTs but are not sufficient to identify all potential sources of bias in a quantitative synthesis. Researchers should also be alert to the possibility and implication of small-study effects for both RCTs and NRS. Small-study effects refers to the generic phenomenon that smaller studies show different, possibly larger, treatment effects than large studies; this may reflect that, there is a higher chance for a small study with positive results (strong treatment effect) to be published compared with a study of a similar size but with negative results (publication bias)[42] or when small studies are of low quality (eg, when at increased risk of outcome selection or reporting bias or due to increased clinical heterogeneity).[43] It is likely that the susceptibility to small-study effects differs between RCTs and NRS in line with differences in the standards that typically govern their design, conduct and reporting; for example, NRS may be potentially at a higher risk of publication bias compared with RCTs. However, these differences may become less of an issue given the recent efforts to improve the design and reporting of NRS. Since small-study effects may affect the validity of meta-analysis results (especially if random-effects model is applied), an evaluation is recommended to determine whether study results are associated with the size of the study. This should be done separately for RCTs and for NRS, and if possible, also separately for different types of NRS. This assessment can, for example, be based on a funnel plots of study results.[44] Unfortunately, statistical tests for analysing funnel plots suffer from low power and cannot determine definitively whether meta-analysis results are invalid.[45 46] Accordingly, their use is best limited to exploring (rather than trying to confirm) any concerns about publication bias.

Steps of a quantitative analysis of effect estimates across study designs (5–7)

Meta-analysis, the statistical technique to combine the study results into a weighted average, while accounting for the precision of each study estimate, is widely being employed by decisionmakers to quantitatively synthesise evidence from multiple sources ('totality of evidence approach') and produce comparative estimates of effects for the new technology under assessment compared with standard clinical care. Researchers and healthcare decision-makers should consider the following underlying concerns about NRS before combining results from such studies with RCT data in evidence synthesis:

- NRS are more prone to selection and confounding biases than RCTs.
- Estimands defined in RCTs are not necessarily transferrable towards NRS and vice versa. It is, therefore, important to consider the applicability of study results with respect to the review question.

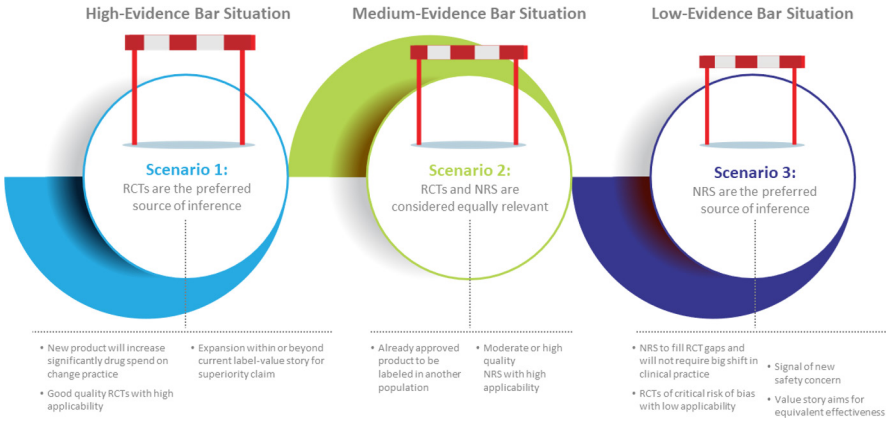


Figure 2 Evidence generation needs in healthcare decision setting and use of non-randomised studies (NRS) with randomised controlled trials (RCTs).

- Special consideration needs to be paid in selecting the appropriate techniques for dealing with incorrect or missing values (including outcomes).
- Analyses that weight studies by simple quality scores should be avoided.
- Summary effect estimates of treatments that are based on data from RCTs and NRS, may be biased and imprecise, even after applying the recommended statistical methodology. Further sensitivity analyses are always recommended to explore the impact of modelling assumptions.
- When RCTs and NRS are combined through network meta-analysis, there is a need for deeper investigation of 'transitivity' (ie, no systematic differences between the available treatment comparisons other than the treatments being compared in the analysis) than when only RCTs are included.

Step 5: selecting the most relevant analytical scenario

The critical appraisal tools cited in step 3 along the other critical domains identified by this group which are not covered by the existing tools may help identify which NRS have enough validity to be considered for evidence synthesis along RCTs. However, given that these tools primarily aim to assess the internal validity of studies, researchers are urged to also consider issues around external validity (generalisability or applicability) in relation to the PICOST criteria set up for the specific research question under assessment. It is not advisable to use NRS which are assessed at critical risk of bias (step 3) for combined analysis with RCTs to avoid misleading and untrustworthy meta-analysed results. This approach differs from that recommended in RCTs meta-analyses, where low-quality studies are usually only excluded in a sensitivity analysis. Depending on the context of the review, the research question and the contribution of NRS in the healthcare decision making problem (eg, if the product is for primary or secondary approval), it may be necessary to perform a critical appraisal separately for each outcome. For example, the presence of selection bias may be less relevant when assessing safety outcomes as compared with effectiveness outcomes.

We consider below three analytical scenarios that may generate new evidence and various examples of weighting between new (RCT) and prior (NRS) evidence for an effectiveness labelling change or an assessment of new products (figures 2 and 3). The selection of the most applicable scenario for a given healthcare decision problem will depend on the (1) clinical context ('relevance or applicability'), (2) completeness of RCT data ('evidence gaps') and (3) the magnitude and direction of possible biases of NRS ('data rigour or quality'). These scenarios are linked with the hypothetical examples of the types of studies (RCTs, NRS) that may be primarily considered for regulatory decision- making as detailed in the white paper by Duke Margolis Center for Health Policy.[17] The corresponding methods outlined in this framework

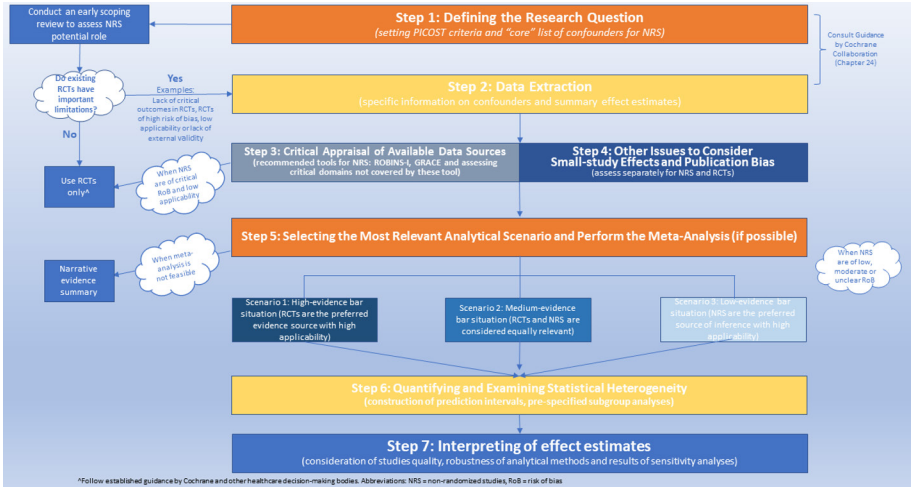


Figure 3 A seven-sStep decision algorithm for the synthesis of non-randomised studies (NRS) and randomised controlled trials (RCTs) in healthcare decision-making (ISPE CER SIG framework). CER, comparative effectiveness research; ISPE, International Society for Pharmacoepidemiology; PICOST, Population, Intervention, Comparison, Outcomes, Study design and Time; SIG, Special Interest Group.

are based on generalised linear mixed models and can be used to summarise many types of association (eg, HR, OR and change score). A critical consideration for each of these scenarios is the attempt to quantify bias in NRS. A helpful review of methods and results from previous studies comparing (and sometimes meta-analysing) RCTs with NRS is provided in the HTA No. 21.7 by the UK National Institute for Health and Research online supplemental appendix 2.[47]

For studies assessed as 'unclear risk of bias', their a priori exclusion from further analyses with RCTs is not recommended. However, for scenarios 2 and 3, their inclusion may directly affect the pooled treatment effect estimates (in comparison to scenario 1) and should therefore only be explored in a sensitivity analysis. More specifically, any bias concerns about treatment effects estimates should be explored at a later step using predesigned sensitivity analyses.

Before any cross-design synthesis of RCTs and NRS is considered, the direction of treatment effects between study designs should be investigated and assessed if it differs substantially (eg, evidence from NRS suggests no effect whereas good-quality RCTs suggest a strong effect). Several reviews have found little difference between the evidence from observational studies and RCTs,[48–50] but counterexamples exist.[51]

Furthermore, it is important that appropriate statistical models are applied to combine comparative treatment effects from NRS and RCTs, as studies will often differ with respect to their validity (risk of bias) and applicability.[52] It is rarely justifiable to directly combine the effect estimates from all the randomised and non-randomised evidence in a meta-analysis without any type of cross-design statistical adjustment.[53 54] In many situations, the observed differences between the results from RCTs and NRS are prone to much uncertainty. It is therefore recommended to adopt analytical methods that distinguish between the two data sources (RCTs and NRS) and allow for some bias corrections (when this discrepancy cannot be explained by differences in study design and selection of populations). The implementation of these methods is not straightforward and will often require advanced statistical support. A description of approaches for combining RCTs with NRS has been presented as part of the GetReal WorkPackage 4 and is summarised in online supplemental table 2.

Scenario 1: high-evidence bar situation

RCTs are generally considered the gold standard for generating evidence about the efficacy of medical interventions as they are designed to test treatment effects while essentially balancing for all other factors (known and unknown) that may affect their response to treatment. For some decision making problems, such as a new product likely to significantly increase drug spending or a product label expansion supporting a superiority claim, the evidence generation needs are high and RCTs are the preferred source of estimating comparative treatment effect estimates. However, in some circumstances described previously, there may still be a desire to augment the evidence from RCTs with results from NRS without directly performing cross-study design meta-analysis. This strategy may be instrumental when RCT evidence is very imprecise (eg, results are only reported for surrogate outcomes), not reflective of the patient population of interest or not covering important patient groups, even when the evidentiary needs for the decision problem are high.

In this circumstance, a natural approach is to treat the NRS data as prior evidence for the RCT analyses, adopting a Bayesian estimation framework.[54 55] Here, the NRS data are summarised using a (network) meta-analysis and, if necessary, adjusted for a predefined amount of bias. The bias adjustment can be performed in different ways depending on the source of bias and the granularity of available data. For instance, it is possible to directly adjust for (differences in) measurement error or missing data with imputation methods. Alternatively, it is possible to apply corrections to the study results by eliciting expert opinion or using the credibility ceiling correction. The latter approach (credibility ceiling correction) assumes that no single NRS can provide a maximum credibility ceiling above a certain percentage.[56] The results of the NRS analysis are then used as the prior distribution for the (network) meta-analysis of the RCT data. In other words, this approach will 'pull' the treatment-effect estimates from the RCTs toward the (adjusted) summary effects from the NRS. By default, the prior distribution(s) has the precision of the summary effect estimate(s) from the NRS. However, it is possible to decrease the precision of the prior distribution(s) by considering additional sources of uncertainty, such as the presence of between-study heterogeneity in the NRS results. A sensitivity analysis should also be conducted to adjust each NRS in the meta-analysis for various ceiling percentages and to observe the direction of effects and consistency in the conclusions obtained (step 6).

Scenario 2: medium-evidence bar situation

In some circumstances, NRS are likely to provide additional (complementary) information about the effectiveness and safety of medical interventions, but their results cannot be directly used as prior information for the RCT results. This situation may arise when RCTs only provide evidence on short-term or surrogate endpoints (eg, when RCTs have low applicability),[57] or when an approved product is being tested in another (beyond its marketing authorisation) indication. Treatment effects are then likely to differ between the RCTs and the NRS, such that greater efforts are needed to disentangle the potential sources of between-study design heterogeneity.

A simple solution is to consider the use of three-level hierarchical models.[54 55] These regression-based models use the first level to model variation within individual studies, the second level to model variation between studies, and the third level to model variation between RCTs and NRS.[58] They typically assume that the treatment effects are different, but exchangeable, across different types of studies, and allow for differences in between-study heterogeneity within randomised and non-randomised data sources.

Like traditional meta-analysis methods, summary estimates of treatment effects generated by three-level hierarchical models represent a weighted average of the included studies. However, the meta-analysis now yields a summary of treatment effect for each distinct study design and an overall treatment effect across all study designs. The overall treatment effect is then pulled towards the results from large, homogeneous studies that share a common design. In addition, because the contribution of each study is adjusted for its study design, estimates of precision are likely to better reflect the various sources of uncertainty (due to bias or heterogeneity).

Scenario 3: low-evidence bar situation

In some situations, NRS may be the most reliable source of inference for obtaining and assessing the external validity of comparative effect estimates. It is, for instance, possible that published RCTs are scarce, or have very poor quality. It is also possible that results from RCTs have limited external validity or applicability about the research question, for instance in postmarketing settings

where the focus is on safety and long-term outcomes. Although corresponding pooled results can be summarised using traditional random-effects meta-analysis methods, researchers should always evaluate the impact of potential bias(es) arising from the synthesis of individual NRS alone or in combination with RCTs.[59] Methods to record and assess the types of bias(es) at the NRS (study) level have already been captured under steps 2 and 3.

While developing this framework, the application of several methods was reviewed. These methods have been developed for adjusting for bias in a meta-analysis of NRS and RCTs which may be applicable to different healthcare decision making problems, depending on the specific biases associated with the NRS under consideration. These methodological approaches may adjust the meta-analysis model to account for bias parameters (eg, for ascertainment or disease onset misclassification bias,[60–63] misclassification of exposure or outcome,[64] or uncontrolled confounding).[65] The application of bias adjustments has been widely advocated in the estimation of treatment effects by NRS and should also be considered during their meta-analysis.[66–69] For healthcare decision-makers, quantifying bias is a critical step, for instance, through clinician and patient surveys or consensus meetings.[67] This method proposes to construct an idealised study (where all questions can be answered) and ask assessors to elicit the likely magnitude and variance of various types of biases including both internal and external validity bias.[70] Expert elicitation is a complex task, because the magnitude of bias always remains uncertain and quantifying the level of uncertainty is part of the elicitation process. Estimates of bias(es) can then be used to adjust the extracted treatmente ffect estimates, and/or to decrease the precision of NRS results accounting for both the magnitude and the uncertainty of the potential bias(es). The adjusted estimates can then be pooled using traditional (network) meta-analysis methods—an approach known as a design-adjusted analysis.[55] This method, which aims to reduce decision uncertainty, is widely used in HTAs, particularly for economic modelling.[70]

Alternatively, it is possible to perform data-driven bias adjustments in evidence synthesis. Several methods have been proposed for integrating bias modelling in the meta-analysis, and these commonly assume that (some of) the NRS overestimate the true treatment effect.[58] A recent approach called hierarchical meta-regression (HMR) distinguishes between biased and unbiased study results and derives the risk of bias automatically from observed study design features (eg, the results from an appraisal of study's quality[71 72]). A mixture model is then used to convert the observed treatment-effect estimate into an unbiased effect. Thus, HMR can identify studies presenting conflicting evidence and downplay their contribution in the (network) meta-analysis.

Step 6: quantifying and examining statistical heterogeneity

As previously mentioned, it will often be difficult to avoid statistical heterogeneity in a meta-analysis especially when NRS are included. These studies are often prone to residual confounding and may therefore affect pooled estimates of relative treatment effects even when excluding studies at high risk of bias and/or adopting advanced meta-analysis methods. Therefore, exploring differences between RCTs and NRS results in a meta-analysis is an important step in evidence synthesis.[73–76] This can be achieved by adopting random effects models and quantifying the presence of between-study heterogeneity. In practice, when substantial between-study heterogeneity is present, the 'average' effect may no longer be an appropriate summary estimate. Between-study heterogeneity typically occurs when there are interactions between the treatment effect and the study or a study-level variable, or when the treatment effect varies across patients.[77] To assist the interpretation of between-study heterogeneity, researchers may derive τ^2 or I^2 statistics although these metrics have limited clinical interpretation, especially when used in isolation. More relevant for healthcare decision making is the construction of (approximate) prediction intervals. Prediction interval depict the expected range of true effects in future studies if those settings are similar to the settings included in the meta-analysis (please see further details on how to calculate a prediction interval in the publication by Riley *et al*[78]) which offer advantages in examining whether the variation of effect estimates is attributable to between-study heterogeneity and enabling the decision makers to interpret the impact of heterogeneity in relation to harm and clinical benefit thresholds (commonly used by decision-makers).[78 79] Meta-regression might be also a way of exploring potential sources of between-study heterogeneity, such as the presence of publication bias, differences in study design or differences in the control treatment.[80] However, this approach has very low power and is prone to ecological bias when used to investigate summarised participant-level characteristics (eg, mean age) as modifier of treatment effect. Several authors have, therefore, recommended the retrieval and inclusion of individual-participant data,[81] a topic beyond the scope of this manuscript.

Finally, in all analytical scenarios, as previously noted, prespecified sensitivity analyses should be performed to assess the extent to which the cross-synthesis results from NRS and RCTs are credible and, understanding the impact of assumptions made in the selection and analysis of NRS by omitting individual studies (eg, in terms of NRS study design, study quality, outcomes time points or other statistical methods employed) on the treatment effects. These sensitivity analyses should focus on key issues that may potentially introduce uncertainty in the estimates of effects (even though it might be, in some cases, difficult to quantify) and lower the credibility of NRS in the decision making.

Step 7: interpretation of effect estimates

Aiming towards increasing the credibility of treatment effects estimates by inclusion of NRS, the interpretation of the results of any quantitative synthesis of NRS and RCTs should always consider the following three points: (1) the quality of the included studies (both RCTs and NRS), (2) the robustness of adopted analytical methods and (3) the results of any sensitivity analyses. Since random-effects summary estimates may be of limited value in the presence of substantial heterogeneity, prediction intervals may help to explore their potential impact on decision making (although it can only be calculated when the meta-analysis includes at least three studies and is most appropriate when the studies have low risk of bias).[78] This group discussed how this step of the framework is heavily dependent on the methods, and the context stipulated different regulatory, payer or reimbursement bodies and the level of certainty/confidence in results they set as thresholds in their decision making.[82] For example, there may be a preference for certain types of evidence (including RCT and NRS) to support economic arguments in the postregulatory environment. Furthermore, when a health economic analysis for new medical technologies is required in the technology's assessment, a probabilistic scenario analysis of economic modelling can provide different thresholds of 'trust' in the results generated by combining NRS and RCT data. However, for organisations that only assess the clinical effects of new products, more scrutiny may be placed on the selection of the most appropriate comparative analytical approach and the consistency of results between NRS and RCTs.

Conclusion

Recent developments in the NRS landscape and the lack of trust among stakeholders in the wider application of such evidence in healthcare decision making have highlighted the pressing need for methodological standards in this area. In particular, this requires widespread understanding of, and familiarity with, the methodological and analytical approaches of NRS that are most likely to offer decision making bodies the level of scientific rigour and certainty they require to rely on evidence from NRS when combined with RCTs. There must also be a recognition of key challenges in the use and interpretation of NRS in this setting and the fact that these will vary with the specific methodological or clinical issues to the decision problem under consideration. Advanced statistical support may be required to undertake some of the proposed analyses of combined analyses of RCTs and NRS. Against this background, our proposed framework aims to set up clear guidance for considering evidence across different study designs—specifically RCTs and NRS—and ensure appropriate, well-established approaches are followed in combining evidence from these sources. We believe it will improve transparency and build confidence in the use of NRS effect estimates and will prompt discussion among regulators and healthcare decision makers who may be sceptical toward the standardised adoption of these novel methodologies (previously described as the 'methodology aversion in drug regulation').[83] The timing of this framework development is also highly relevant, given that many decision making frameworks are currently undergoing revisions to acknowledge and identify ways to incorporate the potential value of NRS in their assessments. However, persistent issues related to poorly reported publications, data inaccessibility from RWE repositories and data governance (which were beyond the scope of this framework) are critical to overcome in order for industry, healthcare bodies and decision makers to explore the added value of NRS and test the application of the proposed methods for our framework. A mandatory national registry for NRS along with strict protocols in analysis and reporting of data (as previously recommended by ISPOR/ISPE taskforce) would provide a platform to further increase the credibility of evidence from NRS.

Therefore, readers are encouraged to consider these recommendations alongside previous guidance related to the design of NRS such as study registration (particularly for hypothesis-evaluating treatment effectiveness studies), data collection (primary or secondary), source validation and results reproducibility, topics not covered by our framework.[45 53 67 84–88] In the future, expanding this framework by considering analyses involving reweighting RCT evidence with real-world NRS evidence[89] or using individual patient data or syntheses of RCT and NRS to inform the design of subsequent RCTs in a clinical development programme[90] could provide greater clarity in other healthcare situations. Further research on analytical methods that may reduce areas of uncertainty in estimating treatment effects from NRS (such as estimating the degree of error in the estimates, investigating the role of machine learning for improving confounder adjustment in EHRs) is much needed.

The next phase of this framework will be testing and validating the proposed recommendations using case studies from NRS and RCTs in a specific healthcare decision problem and disseminating the findings to a wider audience. This validation stage should provide additional insights into the utility of the framework in a real-world healthcare decision-making setting and, therefore, could be updated with new methodologies and help to build trust in its reproducibility. We hope that our framework may also guide researchers to appropriately design and primarily analyse evidence from NRS (accounting for different types of biases) to meet the high standards rightly expected by healthcare decision-makers and highly deserved by patients.

Author affiliations
[1]Real World Evidence Sciences, Visible Analytics Ltd, Oxford, UK
[2]Division of Pharmacoepidemiology and Pharmacoeconomics, Dept. of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[3]Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, Ontario, Canada
[4]Department of Pharmacy Practice & Administrative Sciences, University of Cincinnati College of Pharmacy, Cincinnati, Ohio, USA
[5]Takeda, Cambridge, Massachusetts, USA
[6]Pharmacy Practice, College of Pharmacy, University of Rhode Island, Kingston, Rhode Island, USA
[7]Health Services, Policy, and Practice, Brown University, Providence, Rhode Island, USA
[8]Real-World Solutions, IQVIA, California, Colorado, USA
[9]Epidemiology and Outcomes Research, Research Outcomes Innovations LLC, New York City, New York, USA
[10]GlaxoSmithKline USA, Philadelphia, Pennsylvania, USA
[11]University of Kentucky, Department of Pharmacy Practice and Science, Lexington, Kentucky, USA
[12]NDORMS, Center for Statistics in Medicine, University of Oxford, Oxford, UK
[13]Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine (LSHTM), London, UK
[14]Department of Public Heath, Environments and Society, Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine (LSHTM), London, UK
[15]Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands
[16]Smart Data Analysis and Statistics, Utrecht, The Netherlands

Acknowledgements This manuscript received endorsement from the ISPE Board of Directors. We are grateful to the ISPE members for their participation to the Delphi survey. We thank Professor Jaime Caro and Professor Keith Abrams for providing critical review of this framework. We also acknowledge the support of Elvira D'Andrea and Lydia Vinals in reviewing some of the supporting materials for the development of this manuscript and Colleen Dumont for editorial support.

Collaborators Recommendations from the Working Group of the International Society for Pharmacoepidemiology (ISPE) Comparative Effectiveness Research Special Interest Group for the cross-design synthesis of evidence and endorsed by the ISPE Board of Directors.

Contributors All authors conceived and developed the framework described in this manuscript. GS and TPAD drafted the manuscript and responded to other authors' comments. All authors reviewed, contributed to revisions and approved the final version of the manuscript.

Funding Funding to support this manuscript development was provided by ISPE (https://www.pharmacoepi.org/ISPE/assets/File/General/FINAL20Call20for20Manuscripts204-24-19.pdf).

Disclaimer This article reflects the views and opinions of the authors and does not necessarily represent the views of the organisations where they are employed.

Competing interests We have read and understood BMJ Evidence-Based Medicine policy on declaration of interests

and declare the following interests: GS is employed by Visible Analytics, Ltd; DB is employed by Takeda; ARZ holds a grant from Sanofi Pasteur (direct to institution); MP owns stocks from Merck, Sanofi, and Johnson & Johnson; MG is employed by GSK; and TD is an advisor to pharma industry.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement N/A.

ORCID iDs
Grammati Sarri http://orcid.org/0000-0001-5536-8038
Elisabetta Patorno http://orcid.org/0000-0002-8809-9898
Dimitri Bennett http://orcid.org/0000-0002-8387-9342
Andrew R Zullo http://orcid.org/0000-0003-1673-4570
Thomas P A Debray http://orcid.org/0000-0002-1790-2719

References

1 **Krause JH**, Saver RS. Real-World evidence in the real world: beyond the FDA. *Am J Law Med* 2018;**44**:161–79.
2 **Duke University Margolis Center for Health Policy**. A framework for regulatory use of real-world evidence, 2017. Available: https://healthpolicy.duke.edu/sites/default/files/atoms/files/rwe_white_paper_2017.09.06.pdf
3 **European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCEPP)**. ENCEPP Considerations on the Definition of Non-Interventional Trials under the Current Legislative Framework ("Clinical Trials Directive" 2001/20/Ec), 2011. Available: http://www.encepp.eu/publications/documents/ENCePPinterpretationofnoninterventionalstudies.pdf
4 **Eichler H-G**, Abadie E, Breckenridge A, *et al*. Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov* 2011;**10**:495–506.
5 **Canadian Agency for Drugs and Technologies in Health**. *Use of real-world evidence in single drug technology assessment processes by health technology assessment and regulatory organizations*. CADTH, 2018.
6 **Food and Drug Administration (FDA)**. *Framework for FDA's Real-World Evidence Program*. FDA, 2018.
7 **Institute for Clinical and Economic Review (ICER)**. *Real world evidence for coverage decisions: opportunities and challenges*, 2018.
8 **Doupi P**, Klemp M, Goettsch W. *Patient registries as instruments for HTa outcomes research: a European perspective*. Value & Outcomes Spotlight, 2016.
9 **European Commission**. eHealth action plan 2012-2020: innovative healthcare for the 21st century, 2012. Available: https://ec.europa.eu/digital-single-market/en/news/ehealth-action-plan-2012-2020-innovative-healthcare-21st-century
10 **Makady A**, de Boer A, Hillege H, *et al*. What is real-world data? A review of definitions based on literature and Stakeholder interviews. *Value Health* 2017;**20**:858–65.
11 **Oehrlein EM**, Graff JS, Harris J, *et al*. Patient-Community perspectives on real-world evidence: enhancing engagement, understanding, and trust. *Patient* 2019;**12**:375–81.
12 **Yuan H**, Ali MS, Brouwer ES, *et al*. Real-World evidence: what it is and what it can tell us according to the International Society for pharmacoepidemiology (IspE) comparative effectiveness research (CER) special interest group (SIG). *Clin Pharmacol Ther* 2018;**104**:239–41.
13 **Duke Margolis Center for Health Policy**. *Adding real-world evidence to a Totality of evidence approach for evaluating marketed product effectiveness*, 2019.
14 **Duke Margolis Center for Health Policy**. *Understanding the need for Non-randomized studies using secondary data to generate real-world evidence for regulatory decision making and demonstrating their credibility*, 2019.
15 **NICE DSU**. *The use of real world data for the estimation of treatment effects in NICE decision making*, 2016.
16 **Chan K**, Nam S, Evans B, *et al*. Developing a framework to incorporate real-world evidence in cancer drug funding decisions: the Canadian real-world evidence for value of cancer drugs (CanREValue) collaboration. *BMJ Open* 2020;**10**:e032884.
17 **Duke University Margolis Center for Health Policy**. Adding real-world evidence to a Totality of evidence approach for evaluating marketed product effectiveness, 2019. Available: https://healthpolicy.duke.edu/publications/adding-real-world-evidence-totality-evidence-approach-evaluating-marketed-product
18 **Higgins JPT TJ**, Chandler J, Cumpston M, *et al*. *Cochrane Handbook for systematic reviews of interventions version 6.0 (updated July 2019)*. Cochrane Collaboration, 2009.
19 **Moher D**, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009;**62**:1006–12.
20 **Vandenbroucke JP**, von Elm E, Altman DG, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;**4**:e297.
21 **Bind M-AC**, Rubin DB. Bridging observational studies and randomized experiments by embedding the former in the latter. *Stat Methods Med Res* 2019;**28**:1958–78.
22 **Greenland S**, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
23 **Ades AE**, Caldwell DM, Reken S, *et al*. Evidence synthesis for decision making 7: a reviewer's checklist. *Med Decis Making* 2013;**33**:679–91.
24 **Lefebvre CGJ**, Briscoe S, Littlewood A, *et al*. Searching for and selecting studies. In: *Cochrane Handbook for systematic reviews of interventions*, 2019.
25 **Mueller M**, D'Addario M, Egger M, *et al*. Methods to systematically review and meta-analyse observational studies: a systematic scoping review of recommendations. *BMC Med Res Methodol* 2018;**18**:44.
26 **Fu R**, Vandermeer BW, Shamliyan TA, *et al*. *Handling continuous outcomes in quantitative synthesis. methods guide for effectiveness and comparative effectiveness reviews*. Rockville (MD): AHRQ Methods for Effective Health Care, 2008.
27 **Booth A**, Clarke M, Ghersi D, *et al*. Establishing a minimum dataset for prospective registration of systematic reviews: an international consultation. *PLoS One* 2011;**6**:e27319.
28 **Wang SV**, Schneeweiss S, Berger ML, *et al*. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies v1.0. *Pharmacoepidemiol Drug Saf* 2017;**26**:1018–32.
29 **Higgins JPT GS**. Chapter 7: Selecting studies and collecting data. In: *Cochrane Handbook for systematic reviews of interventions*. The Cochrane Collaboration, 2011.
30 **Liu Z**, Rich B, Hanley JA. Recovering the RAW data behind a non-parametric survival curve. *Syst Rev* 2014;**3**:151.

31 **Parmar MK**, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;**17**:2815–34.
32 **Tierney JF**, Stewart LA, Ghersi D, *et al*. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;**8**:16.
33 **Austin PC**. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;**46**:399–424.
34 **Martens EP**, Pestman WR, de Boer A, *et al*. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. *Int J Epidemiol* 2008;**37**:1142–7.
35 **Burgess S**. Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Commun Stat Theory Methods* 2017;**46**:786–804.
36 **Pirracchio R**, Carone M, Rigon MR, *et al*. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Stat Methods Med Res* 2016;**25**:1938–54.
37 **Stürmer T**, Wang T, Golightly YM, *et al*. Methodological considerations when analysing and interpreting real-world data. *Rheumatology* 2020;**59**:14–25.
38 **Wells GA**, Shea B, Higgins JP, *et al*. Checklists of methodological issues for review authors to consider when including non-randomized studies in systematic reviews. *Res Synth Methods* 2013;**4**:63–77.
39 **Morton SC**, Costlow MR, Graff JS, *et al*. Standards and guidelines for observational studies: quality is in the eye of the beholder. *J Clin Epidemiol* 2016;**71**:3–10.
40 **Quigley JM**, Thompson JC, Halfpenny NJ, *et al*. Critical appraisal of nonrandomized studies-A review of recommended and commonly used tools. *J Eval Clin Pract* 2019;**25**:44–52.
41 **Cochrane Methods Bias**. Rob 2: a revised Cochrane risk-of-bias tool for randomized trials. Available: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials
42 **Duval S**, Tweedie R, Rothstein H, *et al*. *Publication bias in meta-analysis: prevention, assessment and adjustments*, 2005.
43 **Ioannidis JPA**. Why most published research findings are false. *PLoS Med* 2005;**2**:e124.
44 **Rücker G**, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biom J* 2011;**53**:351–68.
45 **Cox E**, Martin BC, Van Staa T, *et al*. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. *Value Health* 2009;**12**:1053–61.
46 **Johnson ML**, Crown W, Martin BC, *et al*. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. *Value Health* 2009;**12**:1062–73.
47 **Hettle R**, Corbett M, Hinde S, *et al*. The assessment and appraisal of regenerative medicines and cell therapy products: an exploration of methods for review, economic evaluation and appraisal. *Health Technol Assess* 2017;**21**:1–204.
48 **Anglemyer A**, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev* 2014;**4**:MR000034.
49 **Schnell-Inderst P**, Iglesias CP, Arvandi M, *et al*. A bias-adjusted evidence synthesis of RCT and observational data: the case of total hip replacement. *Health Econ* 2017;**26 Suppl 1**:46–69.
50 **Dahabreh IJ**, Sheldrick RC, Paulus JK, *et al*. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *Eur Heart J* 2012;**33**:1893–901.
51 **Hue TF**, Cummings SR, Cauley JA, *et al*. Effect of bisphosphonate use on risk of postmenopausal breast cancer: results from the randomized clinical trials of alendronate and zoledronic acid. *JAMA Intern Med* 2014;**174**:1550–7.
52 **Cave A**, Kurz X, Arlett P. Real-World data for regulatory decision making: challenges and possible solutions for Europe. *Clin Pharmacol Ther* 2019;**106**:36–9.
53 **Public Policy Committee, International Society of Pharmacoepidemiology**. Guidelines for good pharmacoepidemiology practice (Gpp). *Pharmacoepidemiol Drug Saf* 2016;**25**:2–10.
54 **Schmitz S**, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. *Stat Med* 2013;**32**:2935–49.
55 **Efthimiou O**, Mavridis D, Debray TPA, *et al*. Combining randomized and non-randomized evidence in network meta-analysis. *Stat Med* 2017;**36**:1210–26.
56 **Ioannidis JPA**. Commentary: adjusting for bias: a user's guide to performing plastic surgery on meta-analyses of observational studies. *Int J Epidemiol* 2011;**40**:777–9.
57 **Schünemann HJ**, Tugwell P, Reeves BC, *et al*. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods* 2013;**4**:49–62.
58 **Verde PE**, Ohmann C. Combining randomized and non-randomized evidence in clinical research: a review of methods and applications. *Res Synth Methods* 2015;**6**:45–62.
59 **Reeves BC DJ**, Higgins JPT, Shea B. Including non-randomized studies on intervention effects. In: *Cochrane Handbook for systematic reviews of interventions*, 2019.
60 **Greenland S**. Multiple-bias modelling for analysis of observational data (with discussion). *J R Stat Soc Ser A Stat Soc* 2005;**168**:267–306.
61 **Phillippo DM**, Dias S, Welton NJ, *et al*. Threshold analysis as an alternative to grade for assessing confidence in guideline recommendations based on network meta-analyses. *Ann Intern Med* 2019;**170**:538–46.
62 **Thompson S**, Ekelund U, Jebb S, *et al*. A proposed method of bias adjustment for meta-analyses of published observational studies. *Int J Epidemiol* 2011;**40**:765–77.
63 **Wolpert RL**, Mengersen KL. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statist Sci* 2004;**19**:450–71.
64 **Col NF**, Kim JA, Chlebowski RT. Menopausal hormone therapy after breast cancer: a meta-analysis and critical appraisal of the evidence. *Breast Cancer Res* 2005;**7**:R535–40.
65 **Col NF**, Pauker SG. The discrepancy between observational studies and randomized trials of menopausal hormone therapy: did expectations shape experience? *Ann Intern Med* 2003;**139**:923–9.
66 **Lash TL**, Fox MP, MacLehose RF, *et al*. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;**43**:1969–85.
67 **Food and Drug Administration (FDA)**. Best practices for conducting and reporting Pharmacoepidemiologic safety studies using electronic healthcare data. Available: https://www.fda.gov/downloads/drugs/guidances/ucm243537.pdf
68 **The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP)**. *Guide on methodological standards in pharmacoepidemiology (revision 2)*, 2013.
69 **Spiegelhalter DJ**, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Stat Med* 2003;**22**:3687–709.
70 **Turner RM**, Spiegelhalter DJ, Smith GCS, *et al*. Bias modelling in evidence synthesis. *J R Stat Soc Ser A Stat Soc* 2009;**172**:21–47.
71 **Verde PE**. The hierarchical metaregression approach and learning from clinical evidence. *Biom J* 2019;**61**:535–57.
72 **Verde PE**. Two examples of Bayesian evidence synthesis with the hierarchical meta-regression approach, 2017. Available: https://www.intechopen.com/books/bayesian-inference/two-examples-of-bayesian-evidence-synthesis-with-the-hierarchical-meta-regression-approach
73 **Higgins JPT**. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008;**37**:1158–60.
74 **Higgins JPT**, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
75 **Ioannidis JPA**, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ* 2007;**335**:914–6.

76 **Melsen WG**, Bootsma MCJ, Rovers MM, *et al*. The effects of clinical and statistical heterogeneity on the predictive values of results from meta-analyses. *Clin Microbiol Infect* 2014;**20**:123–9.

77 **Dias SSA**, Welton NJ, Ades AE. *Heterogeneity: subgroups, meta-regression, bias and bias-adjustment*. UK: Decision Support Unit (DSU), 2011.

78 **Riley RD**, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;**342**:d549.

79 **IntHout J**, Ioannidis JPA, Rovers MM, *et al*. Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open* 2016;**6**:e010247.

80 **Schmid CH**, Stark PC, Berlin JA, *et al*. Meta-Regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *J Clin Epidemiol* 2004;**57**:683–97.

81 **Debray TP**, Schuit E, Efthimiou O, *et al*. An overview of methods for network meta-analysis using individual participant data: when do benefits arise? *Stat Methods Med Res* 2018;**27**:1351–64.

82 **Nicod E**, Kanavos P. Developing an evidence-based methodological framework to systematically compare HTa coverage decisions: a mixed methods study. *Health Policy* 2016;**120**:35–45.

83 **Bauer P**, König F. The risks of methodology aversion in drug regulation. *Nat Rev Drug Discov* 2014;**13**:317–8.

84 **ISPOR**. Improving transparency in non-interventional research for hypothesis Testing—WHY, what, and how: considerations from the real-world evidence transparency initiative, 2019. Available: https://www.ispor.org/docs/default-source/strategic-initiatives/improving-transparency-in-non-interventional-research-for-hypothesis-testing_final.pdf?sfvrsn=77fb4e97_6

85 **Berger ML**, Martin BC, Husereau D, *et al*. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC good practice Task force report. *Value Health* 2014;**17**:143–56.

86 **Berger ML**, Sox H, Willke RJ, *et al*. Good practices for real-world data studies of treatment and/or comparative effectiveness: recommendations from the joint ISPOR-ISPE special Task force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf* 2017;**26**:1033–9.

87 **eunethta**. Internal Validty of Non-Randomized Studies-(NRS)-on-interventions. Available: http://www.eunethta.eu/outputs/Internal-Validity-of-non-randomized

88 **European Medicines Agency (EMA)**. Guideline on good pharmacovigilance practices (GVP) module VIII-Post-Authorization safety studies (Rev 2), 2016. Available: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf

89 **Happich M**, Brnabic A, Faries D, *et al*. Reweighting Randomized Controlled Trial Evidence to Better Reflect Real Life - A Case Study of the Innovative Medicines Initiative. *Clin Pharmacol Ther* 2020;**108**:817–25.

90 **Martina R**, Jenkins D, Bujkiewicz S, *et al*. The inclusion of real world evidence in clinical development planning. *Trials* 2018;**19**:468.

SUPPLEMENTARY MATERIALS

Appendix 1. Methodology of Framework Development

The ISPE CER SIG working group that developed this framework is composed of 14 members representing different stakeholders (academia, policymakers, pharmaceutical product development, health consultants) covering various geographic jurisdictions. The working group met regularly for more than 12 months and leveraged its expertise to develop the current framework using an iterative process.

The specific objectives of our working group were two-fold:

a) To critically review the existing published evidence covering the following questions:

   1. "How should the quality and compatibility of evidence from NRS and RCTs be assessed, and when is it appropriate to combine evidence from RCTs with NRS?"
   2. "How should NRS and RCT data be combined in a quantitative synthesis to generate reliable comparative effect estimates?"

b) To provide a step-by step guidance for researchers and policymakers when considering the combination of NRS with RCTs to estimate relative effect estimates for healthcare decision-making.

The development of this framework involved a multi-step process, which began with defining the research questions.

In the next step, a combination of methods was applied to address each of the two research objectives. More specifically, supporting information was retrieved through:

- Systematic literature review (SLR) and update (umbrella of SLRs) on tools assessing validity (extent of susceptibility to bias) in NRS (following a review protocol and database searches)

  The systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement [9]. Systematic review protocol and registration are available at https://osf.io/es65q.

Systematic search and eligibility criteria

We searched Pubmed and Embase from inception to November 2019 to identify existing tools that investigated the validity of NRS, specifically case-control and cohort design studies. We excluded guidelines or manuals, tools to review study protocols, tools targeting NRS of non-pharmacological interventions (e.g. surgery) or assessing only one or a few specific types of bias, and tools not available in English language. In parallel, we searched the same electronic databases for systematic reviews of assessment tools of NRS. We then extracted the references of the tools included in the systematic reviews retrieved. We also performed a general search through Google® for grey literature and reviewed any additional information from initiatives, programs or organizations, and suggestions from experts. Full details on the search strategy are reported in the Supplement (Table S1 and S2, online Supplement 1). Three reviewers (E.D., G.S., L.V.) independently removed duplicates and reviewed titles and abstracts of peer-reviewed publications or documents from the grey literature to select eligible tools. Discrepancies were resolved by consensus.

Delphi survey and prespecified framework

Concurrently, we performed a Delphi survey to reach a consensus among content experts about the main methodological challenges (domains) that may threaten the validity of NRS on comparative safety and effectiveness of medications. The survey is available in the online Supplement 2. The panel of experts involved members of the SIG for CER of the ISPE. Detailed information on the Delphi methods and results is reported in Supplementary Figure 1.

Domains and subdomains indicated by the Delphi respondents as major elements that can impact the validity of NRS of medications were used to develop and pilot a framework to evaluate the identified NRS tools. All domains were considered equally important.

Data extraction

Two reviewers (E.D., L.V.) independently extracted general information of the identified tools (first author or name of the tool, year of publication or online availability of the most updated version, type of tool, scope of the tool, non-

randomized study designs evaluated, number of items) and content data related to the prespecified domains of the framework. Discrepancies were resolved by consensus. We categorized the tools as checklists, defined as itemized instruments (including questionnaires) developed to identify the presence or absence of critical elements, or rating scales, defined as itemized instruments aimed to identify the performance of a study at each critical element described in the tool, using a qualitative or quantitative scale.

Data synthesis

General characteristics of the identified tools were summarized with means and standard deviations, for continuous variables, and relative frequencies, for categorical variables. The findings from the online survey and the proportion of tools assessing the prespecified elements of the framework were reported in terms of relative frequencies.

- Identification of publications from previous SLRs to address methodological issues and statistical analysis approaches for combining NRS with RCTs (e.g., Innovative Medicines Initiative [IMI] GetReal, Institute for Clinical and Economic Review [ICER], Duke-Margolis Health Policy Center)
- Pragmatic identification of relevant materials based on the group's knowledge and prior experience (supplementary online searches)

Supplementary Table 1. Search Strategy for the Systematic Literature Review of Quality Tools for NRS

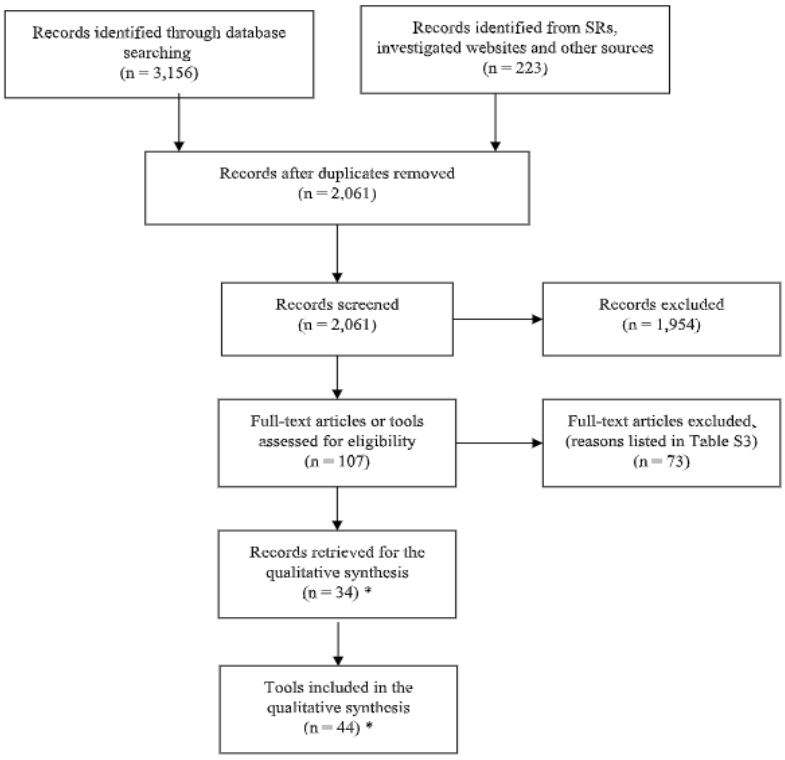
Items	N.	Terms
Critical appraisal tool	#1	"critical" [All Fields] AND "appraisal" [All Fields] AND "tools" [All Fields]
	#2	"critical" [All Fields] AND "appraisal" [All Fields]
	#3	("critical" [All Fields] AND "review" [All Fields]) OR "critical review" [All Fields]) AND form [All Fields]
	#4	("systematic review" [Publication Type] OR "systematic reviews as topic" [MeSH Terms] OR "systematic review"[All Fields]) AND form [All Fields]
	#5	appraisal [All Fields] AND ("research design" [MeSH Terms] OR ("research" [All Fields] AND "design" [All Fields]) OR "research design" [All Fields] OR ("research" [All Fields] AND "methodology" [All Fields]) OR "research methodology"[All Fields])
	#6	("research design" [MeSH Terms] OR ("research" [All Fields] AND "design" [All Fields]) OR "research design" [All Fields]) AND ("review" [Publication Type] OR "review literature as topic" [MeSH Terms] OR "review"[All Fields])
Study reporting tool	#7	"study" [All Fields] AND "reporting" [All Fields] AND "tool" [All Fields]
	#8	"study" [All Fields] AND "reporting" [All Fields]

	#9	"reporting" [All Fields] AND "form" [All Fields] AND ("Studies"[Journal] OR "studies"[All Fields])
	#10	"reporting" [All Fields] AND ("Studies"[Journal] OR "studies"[All Fields])
Tool	#11	"checklist" [MeSH Major Topic] OR "scale*" [Title/Abstract]
	#12	"surveys and questionnaires"[MeSH Major Topic] OR "questionnaire*" [Title/Abstract]
	#13	("tool*" [All Fields] OR "instrument*" [All Fields] OR "checklist*" [All Fields] OR "questionnaire*" [All Fields]) AND ("quality" [All Fields] OR "method*" [All Fields] OR "bias" [All Fields])
Study design	#14	"cohort studies"[MeSH Terms] OR cohort studies [Text Word] OR cohort stud* [All Fields]
	#15	"case-control studies" [MeSH Terms] OR case-control studies [Text Word] OR case control stud* [All Fields]
	#16	Non [All Fields] AND ("random allocation"[MeSH Terms] OR randomized [Text Word]) AND stud* [All Fields]
Systematic review	#17	"systematic review" [Publication Type] OR "systematic reviews as topic"[MeSH Terms] OR "systematic review"[All Fields]
Filters	#18	"humans"[MeSH Terms]
	#19	"Review" [ptyp] OR "systematic" [sb]
Strings		
1st search - tools*	#20	(#1 OR #2 OR #3 OR #4 OR #5 OR #6) AND (OR #14 OR #15 OR #16) AND #18
	#21	(#7 OR #8 OR #9 OR #10) AND OR #14 OR #15 OR #16) AND #18
	#22	(#11 OR #12 OR #13) AND OR #14 OR #15 OR #16) AND #18
2nd search - systematic reviews of tools*	#23	(#1 OR #2 OR #3 OR #4 OR #5 OR #6) AND (OR #14 OR #15 OR #16) AND #18 AND (#17 OR #19)
	#24	(#7 OR #8 OR #9 OR #10) AND OR #14 OR #15 OR #16) AND #18 AND (#17 OR #19)
	#25	(#11 OR #12 OR #13) AND OR #14 OR #15 OR #16) AND #18 AND (#17 OR #19)

*tools are defined as instruments (e.g., qualitative checklists, questionnaires, scoring scales, etc.) that investigate the overall quality of a study, identifying potential biases – either used to critically appraise studies included in a systematic review, or to help in the peer-reviewing process of scientific publications – (i.e., critical appraisal tools), or that support the reporting of research methods and findings (i.e. study reporting tools).
Note: the strings were built in Medline and then adapted to Embase (through Elsevier)

Supplementary Figure 2. PRISMA

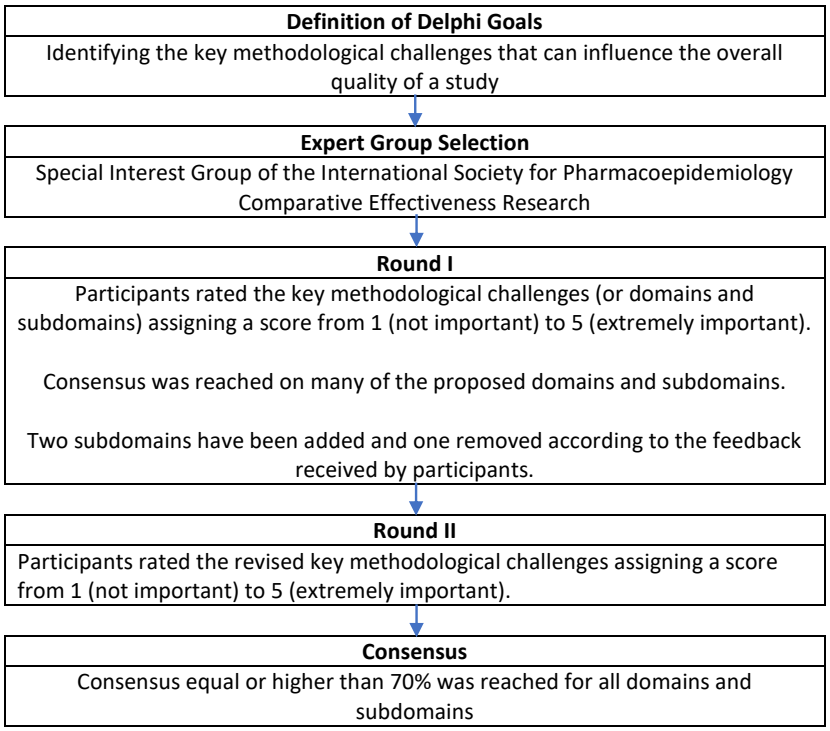
Figure S2. Flowchart of the selection of the tools.



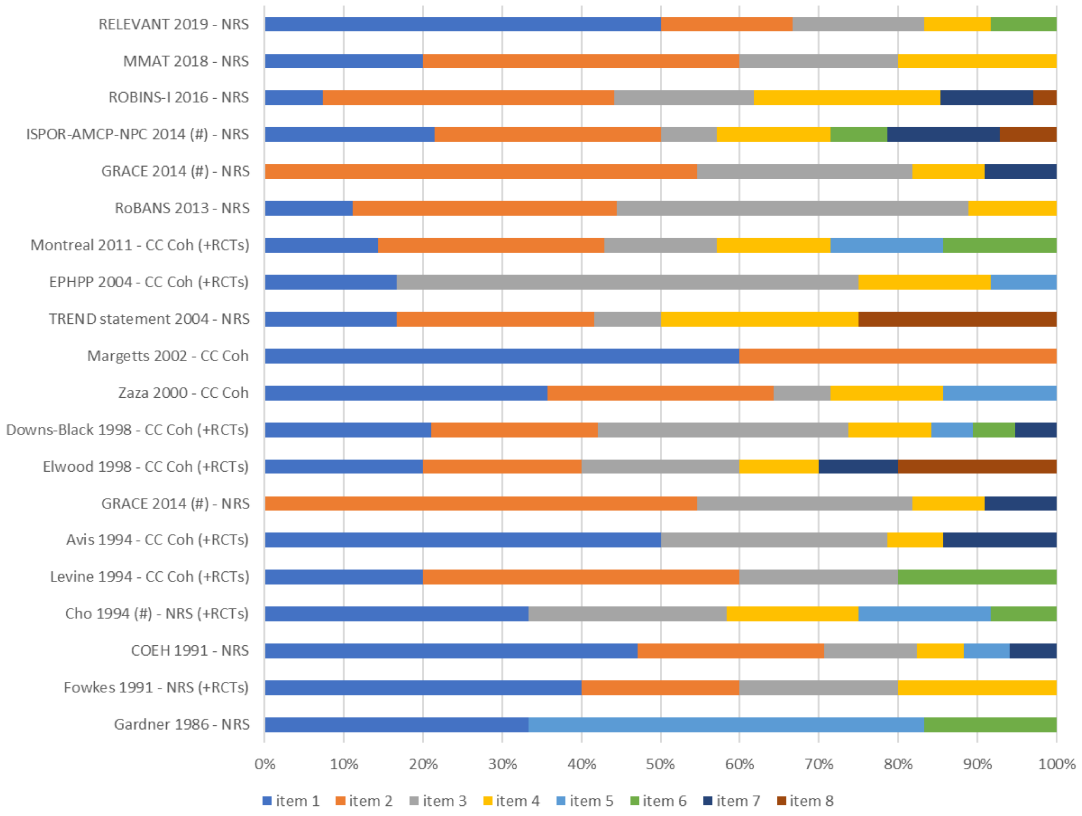
*Ten retrieved articles/records included two different assessment tools for case-control and cohort studies. Thus, the overall number of included tools raised up to 44.

The systematic reviews (SRs) were retrieved through a systematic search. Additional details about the SRs search are available upon request.

Supplementary Figure 1. Flowchart of the Delphi procedure

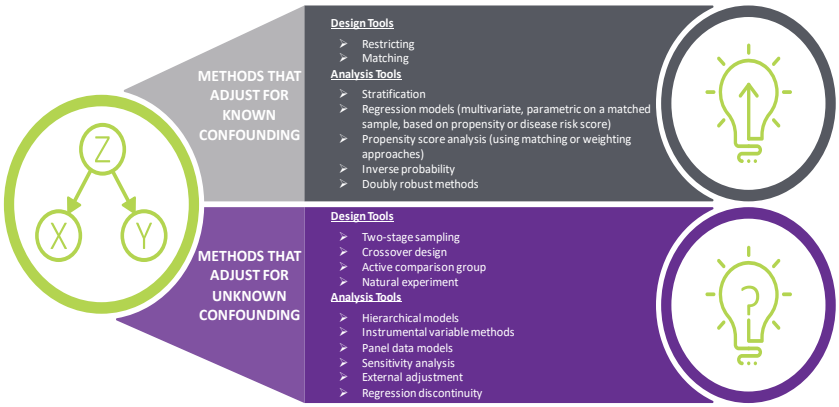


Supplementary Figure 3. Distribution of domains for each reviewed tool of multiple study designs (NRS)



Legend: Item 1 = Methods for selecting participants; item 2 = Definition and measurement of exposure, outcomes, covariates and follow-up; item 3 = Design-specific sources of bias; item 4 = Confounding; item 5 = Lack of appropriateness of statistical analyses; item 6 = Methods for assessing statistical uncertainty in the findings; item 7 = Methods for assessing internal validity; item 8 = Methods for assessing external validity.

Supplementary Figure 4. Summary of Methods to Adjust for Either Known or Unknown Confounding



Sources: RWE navigator (https://rwe-navigator.eu/use-real-world-evidence/adjusting-for-bias-in-non-randomised-and-observational-studies/) and Duke Margolis Center for Health Policy (https://healthpolicy.duke.edu/sites/default/files/atoms/files/non-internvetional_study_credibility_final.pdf)

Supplementary Table 2. Overview of the Presented Approaches for Combining RCTs with NRS Evidence (source: adapted by Efthimiou et al. 2016)

	Using informative priors	Three-level hierarchical models	Bias-adjusted analysis
	Scenario 1	Scenario 2	Scenario 2 and 3
Direct meta-analysis of RCTs and NRS	No	Yes	Yes
Description of the approach	Prior distributions are formulated by meta-analyzing NRS	Data from NRS and RCTs are synthesized separately and then the pooled effect estimates are pooled in a joint meta-analysis.	NRS estimates are adjusted for possible bias and over-precision
How NRS are incorporated	The priors are shifted to account for bias, and/or the variances are inflated to down-weight estimates from NRS. Between study variability (RCTs, NRS) in treatment effect is ignored.	Either NRS can be adjusted separately (according to its features) or adjustment for bias can be performed collectively for each design (on the design-level estimates).	Either NRS can be adjusted separately (according to its features) or common bias parameters can be assumed for all NRS
When to use it (preferably)	When it is infeasible to infer about bias in each study separately	When there are studies pertaining to multiple study designs (RCTs, NRS)	When resources allow inference about bias in each separate NRS

Abbreviations: NRS = non-randomized studies; RCT = randomized controlled trial