Drug Safety (2020) 43:623–633
https://doi.org/10.1007/s40264-020-00944-1

LEADING ARTICLE

A Framework for Methodological Choice and Evidence Assessment for Studies Using External Comparators from Real-World Data

Christen M. Gray[1] · Fiona Grimson[1] · Deborah Layton[1,2,3] · Stuart Pocock[4] · Joseph Kim[1,4,5]

Published online: 21 May 2020

Abstract

Several approaches have been proposed recently to accelerate the pathway from drug discovery to patient access. These include novel designs such as using controls external to the clinical trial where standard randomised controls are not feasible. In parallel, there has been rapid growth in the application of routinely collected healthcare 'real-world' data for post-market safety and effectiveness studies. Thus, using real-world data to establish an external comparator arm in clinical trials is a natural next step. Regulatory authorities have begun to endorse the use of external comparators in certain circumstances, with some positive outcomes for new drug approvals. Given the potential to introduce bias associated with observational studies, there is a need for recommendations on how external comparators should be best used. In this article, we propose an evaluation framework for real-world data external comparator studies that enables full assessment of available evidence and related bias. We define the principle of exchangeability and discuss the applicability of criteria described by Pocock for consideration of the exchangeability of the external and trial populations. We explore how trial designs using real-world data external comparators fit within the evidence hierarchy and propose a four-step process for good conduct of external comparator studies. This process is intended to maximise the quality of evidence based on careful study design and the combination of covariate balancing, bias analysis and combining outcomes.

1 Introduction

Medical research strives to ascertain the efficacy and safety of novel therapeutic substances using the highest quality evidence while addressing the need to bring life-saving therapies to market as quickly as possible. The use of randomised controlled trials (RCTs) has revolutionised the development of medicine in the 20th century, providing the gold standard for generating evidence to assess the efficacy of medicines. Randomised controlled trials may, however, not always be possible especially in severe or rare disease for multiple reasons including slow recruitment. Recently, approaches

Key Points

The strength of the evidence stemming from clinical trials performed with RWD external comparators may be evaluated in terms of study design and exchangeability between the external and trial populations.

Given the challenge of combining observational data from routinely healthcare sources with clinical trial data, additional rigor must be integrated into planning and analytical execution via careful feasibility assessment and quantitative bias analysis.

Bayesian dynamic borrowing methods for combining outcomes from RWD external comparators and the internal comparator arm of an RCT should be considered.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40264-020-00944-1) contains supplementary material, which is available to authorized users.

✉ Christen M. Gray
christen.gray@iqvia.com

1 EMEA Centre of Excellence for Retrospective Studies, IQVIA, London, UK

2 School of Pharmacy and Bioengineering, Keele University, Staffordshire, UK

3 School of Pharmacy and Biomedical Sciences, University of Portsmouth, Portsmouth, UK

4 Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

5 School of Pharmacy, University College London, London, UK

have been developed to accelerate the pathway from drug discovery to patient access. These include revisions to the regulatory framework, such as adaptive pathways [1, 2], and the introduction of flexible designs such as adaptive designs and umbrella trials [3–8].

As the accelerated approval pathway has matured, the number of products receiving marketing approval based on data from non-randomised single-arm trials (SATs) has increased. Between January 1999 and May 2014, the European Medicines Agency (EMA) issued 795 approvals, including 44 solely on evidence from SATs, and the US Food and Drug Administration (FDA) issued 774 and 60 respectively in the same period [9]. The primary disease area where SAT data were used was oncology, with 49/74 indications (66%) being for haematological malignancies or solid tumours. These findings are echoed in the approval of orphan medical products. The European Union approved 125 orphan medical products between 1999 and 2014, one third of which used SAT data and another third did not use randomisation [10].

There is interest in designs that use non-randomised 'control' patients external to the clinical trial for comparison purposes, to strengthen the evidence of a SAT or RCT where patient pools are limited [11, 12]. The external comparator patients are not part of the same trial as those receiving the investigational product. They may be receiving the best standard of care treatment or be untreated, and may be sourced from previous trials, observational studies, registries or databases of routine healthcare. We refer to these patients as 'external comparators', but other synonyms include 'historical controls', 'synthetic controls', 'natural history controls' or 'external controls'. In 2001, European Union guidance on the choice of control groups in clinical trials was published including external comparators, but such a design was regarded at that time as an undesirable option [13]. Reasons included the re-introduction of biases inherent in observational studies that randomisation was intended to mitigate. Nevertheless, external comparators are increasingly used in regulatory decision making. For example, Alecensa (alectinib) received accelerated FDA approval in December 2015, and was conditionally approved by the EMA in February 2017 [14, 15] for the treatment of a specific form of lung cancer. Data come from a SAT with additional evidence of effectiveness relative to standard of care from an external comparator arm. Other examples of EMA approval include Blincyto (blinatumomab) for the treatment of a rare form of leukemia in 2015 [16] and Zalmoxis, an immunogene therapy for high-risk haematological malignancies in 2016 [17].

In addition to innovation in design, there has been rapid growth in the availability and use of electronically captured routine healthcare data (sometimes termed "real-world data" [RWD] [18]), to evaluate post-market safety and effectiveness [19]. It is a natural next step that RWD be proposed for use in establishing an external comparator arm in clinical trials [11, 12]. Real-world data more often reflect the typical use of treatments and tend to encompass patients with widely varying characteristics and co-morbidities than do clinical trials. Therefore, studies using RWD can be more representative of the population requiring treatment in clinical practice (i.e. externally valid). Accordingly, use of RWD in post-authorisation safety studies has become standard practice [20]. Real-world data may contain detailed clinical information; however, the data arise from systems supporting routine clinical practice rather than research. Handling such data requires special considerations.

There is a clear need for recommendations on how RWD external comparators are best used, both through study design and analytical approaches, to maximise the quality of evidence. Fears regarding a lack of predictability in regulatory requirements and rejection of non-standard methods remain real despite publications from regulators such as the EMA concept paper on the extrapolation of safety and efficacy data across populations [21]. Furthermore, the EMA advised in 2006 that historical RWD may be incorporated into the analytical framework through appropriate statistical methods [22]. While the EMA has offered little direct guidance on external comparators since that publication, a report by the Head of Medicines Agency/EMA Joint Big Data Taskforce may provide the greatest insight into current thinking by European regulators [18]. This report provides recommendations regarding the regulatory acceptability of big data, which in this context also relates to individual clinical trials or RWD sources that may be pooled or linked such that the data assume characteristics of big data. In that report, specific actions were noted to improve regulatory guidelines and information for big data. In contrast, the FDA has openly endorsed the use of external comparator studies drawing on RWD in specific circumstances [23].

This article proposes an evaluation framework for RWD external comparator studies that can be applied to assessment of the evidence and related bias. After introducing the concept of exchangeability in Sect. 2, we discuss exchangeability criteria as applied to the internal trial patients and external comparators in Sect. 3. In Sect. 4, we explore how the exchangeability of an external comparator arm with the trial patients can be improved by careful consideration of study design. In Sect. 5, we recommend a four-step approach for analysing a trial with external comparators starting with a feasibility assessment, and we discuss methods available for use in the context of this approach depending upon the external comparator study design. We explore what factors can be satisfied by the study design as well as those that may be controlled by appropriate analytical methods. Finally, in Sect. 6 we discuss steps that might be taken in the future.

2 Exchangeability

The concept of exchangeability relates to how well the unexposed (comparator) group provides an approximation for the disease experience of the exposed group, had they not been exposed. The validity of external comparator patients lies in the exchangeability with the internal trial patients. In this section, we discuss the minimum criteria that need to be met to assume exchangeability exists and explore how RWD external comparators might be expected to depart from those criteria.

In 1976, Pocock described six necessary conditions for the external comparator group to be exchangeable with the randomised internal controls of a clinical trial (Table 1) [24]. The FDA regulators have cited Pocock's exchangeability criteria in their reviews [25]. These criteria provide a reference for assessment of potential sources of bias when considering the use of RWD as a source of external comparators. The more exchangeable the two populations (internal and external), the better. The more potential discordance that can be identified (such that groups may only be partially or non-exchangeable), the lower the strength of the evidence. Key factors include selection criteria, confounders and outcomes; these may encompass both measured and unmeasured variables. In Table 1, the specific biases associated with failure to meet each of Pocock's criteria are listed, together with some study design considerations specific to the use of RWD for external comparators.

Only a randomised internal control group within the clinical trial would allow us to directly evaluate the similarity of the distribution of the outcome(s) and impact of unmeasured variables, which is not possible with a SAT. We discuss the implications of specific study designs in terms of exchangeability in Sect. 3.

3 Trial Study Designs Using Real-World Data External Comparators and Their Place in the Hierarchy of Evidence

In this section, we explore how trial designs using RWD external comparators fit within an evidence hierarchy. Readers may be familiar with the principle of evidence-based medicine and the hierarchy of evidence often displayed as a pyramid, reflecting risk of bias (internal validity) with

Table 1. Pocock's criteria of exchangeable populations, potential bias from lack of exchangeability and considerations when using real-world data (RWD)

Exchangeability criterion	Potential bias if non-exchangeable	Study design considerations for RWD[a]
Subject to the same eligibility criteria	Selection bias/confounding (measured and unmeasured)	The RCT eligibility criteria must be adapted to data that are routinely captured in RWD [12]. Laboratory test intervals will be more irregular. When a record of a co-morbidity is lacking in the database, that co-morbidity will be assumed not to be present in the patient
Distributions of important patient characteristics	Confounding (measured and unmeasured), information bias	All characteristics important to the natural history of the disease should be captured where possible. These may not be recorded in the same manner in both the RCT and RWD
Identical treatment	Positive treatment bias (i.e. placebo effect)	An active treatment comparator is preferable to an untreated comparator, particularly where the trial treatment is invasive (i.e. chemotherapy). RWD may rely assumptions of treatment based on prescription records rather than dispensed treatment
Treatment outcome(s) evaluated in the same manner	Information bias	RWD may need to rely on alternative diagnostics, proxies or passive reporting of outcomes of interest by patient to a healthcare professional
Collected recently	Surveillance bias, unmeasured confounding	Diagnosis of disease, rates of disease, standard of care and reporting of adverse events by patients all change over time
Collected in the same setting, by the same investigators	Unmeasured confounding	RWD will not typically be able to satisfy this criterion, leading to a source of unmeasured confounding

RCT randomised controlled trial

[a] "Real-world data" indicating routinely collected healthcare data

weaker designs at the base and stronger designs at the tip. Analogously, a hierarchical framework can be considered for the quality of evidence offered by RWD external comparator designs (Fig. 1). There are two primary study designs making use of RWD external comparators:

1. Supplemented SAT: A SAT wherein the sole source of controls are RWD external comparators.
2. Augmented RCT: An RCT with reduced sample size in the internal control arm (*N*:1 randomisation ratio), which is then augmented with RWD external comparators.

These can be compared with more traditional trial designs: a SAT (with no source of controls); and a standard RCT with either an equivalent number of patients in each arm (1:1 randomisation ratio) or a ratio that favours the comparator arm (1:*N* ratio).

In addition to these two basic designs, one can combine a fully powered RCT with the use of RWD in a number of ways to gain the advantages of both, i.e. internal and external validity, reduction in loss to follow-up, pathway to evaluating long-term outcomes and potentially establishing reference values for other trials using external comparators [11, 26, 27]. Alternatively, a group of external patients may be used without any direct comparison to create a benchmark or group of natural history comparators.

3.1 Supplemented Single-Arm Trials

In supplemented SATs, external comparators provide the sole source of controls. While supplemented SATs cannot replace RCTs [28], our opinion is that when performing a SAT it is preferable to supplement with an external comparator to inform safety and efficacy than not to have any comparator. While some qualities of exchangeability can be, and should be, assessed for a supplemented SAT, the assessment is limited to variables measured in both the internal trial and the RWD comparator patients, thus only partial exchangeability can be achieved. The greater the proportion of measurable key factors that exist will determine to what extent exchangeability can be assessed and the resulting quality of evidence arising from a trial of this design.

3.2 Augmented Randomised Controlled Trials

In rare circumstances, RCTs may have unequal randomisation ratios, particularly in rare diseases (see Box 1). In these cases, external comparators can be used to augment the RCT by adding power to a smaller internal control group.

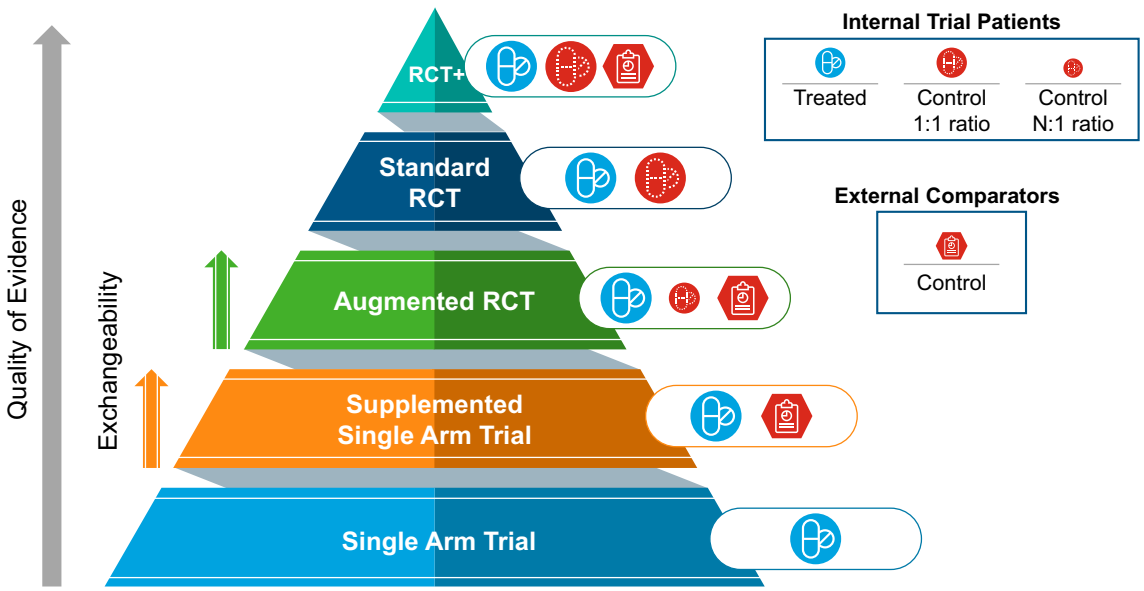

Fig. 1 Study design hierarchy of evidence. The proposed hierarchy of evidence for study designs in the context of use of an external comparator arm from real-world data in comparison to a standard randomised controlled trial (RCT) or a single-arm trial. The quality of evidence, as indicated by the filled arrow, is expected to increase as one goes from a single-arm trial to an RCT; similarly, within the designs for trials using real-world data external comparators, the quality of evidence is dependent upon exchangeability, as indicated by the striped arrows, as it is expected to increase as the exchangeability status between the trial patients and the external comparators transitions from being poor (non-exchangeable), partially exchangeable or completely exchangeable. RCT+ represents study designs that go above and beyond by having a fully powered RCT complemented by external data

Box 1. Reasons for unequal allocation of patients between treatment and control groups

- Safety information – where a larger treatment group provides more power to detect adverse events
- Product information – where a larger treatment group provides more information about a novel therapy
- Recruitment – where subjects are more willing to participate if there is a higher chance of receiving the treatment
- Loss to follow-up – where one arm may have a greater risk of loss to follow-up.
- Cost – where one arm is less expensive than another.

We expect the quality of evidence produced by an augmented RCT to be between a standard RCT and a supplemented SAT. The quality of the evidence depends on the evaluation of the exchangeability between the trial patients and the RWD external comparators, and whether a significant amount of information (i.e. power) is added after appropriately adjusting for exchangeability.

4 Methods

To evaluate exchangeability and potential bias in studies with RWD external comparators, we propose a four-step approach to the design and analysis (Fig. 2). Steps 1–3 are applicable to all trial designs using RWD external comparators, and are reminiscent of the steps for the direct comparison of randomised and observational studies outlined by Lodi et al. building on the work of Hernan and Robins in emulating a hypothetical clinical trial (i.e. a target trial) [29, 30]. We recommend identifying key factors contributing to bias (Step 1), adjusting for baseline characteristics and confounders in analysis (Step 2), as well as modelling and quantifying the potential bias from other key factors, in a process called quantitative bias analysis [QBA] (Step 3). Quantitative bias analysis may be performed at multiple points during the design and analysis of an external comparator study.

If the study is a supplemented SAT without an internal control group, steps 1–3 provide a basis for judging exchangeability and the quality of evidence. Step 4 is applicable in an augmented RCT with an internal control. In Step 4, outcomes from the external comparators and the internal control group are combined in a Bayesian analysis, termed 'dynamic borrowing' [31]. In this method, power from the external comparators is borrowed depending upon the exchangeability with the internal control group.

5 Step 1: Identification of Key Factors Affecting Bias

Step 1 is the initial planning phase, where key factors influencing bias should be identified. The sources of potential bias of highest impact for RWE external comparator studies include selection bias, unmeasured confounding and information bias (i.e. misclassification, measurement and missing data). Unmeasured confounding can arise when there is differential selection not captured in the eligibility criteria or from important patient characteristics not included in the study, e.g. when there are gaps in RWD availability and not all relevant variables are captured [28]. Misclassification and measurement error could result from using different procedures for the measurement of any variable in the RWD vs the trial. The level and extent of missingness are also likely to vary between RWD sources compared to the intensive data collection systems for trials [32–34].

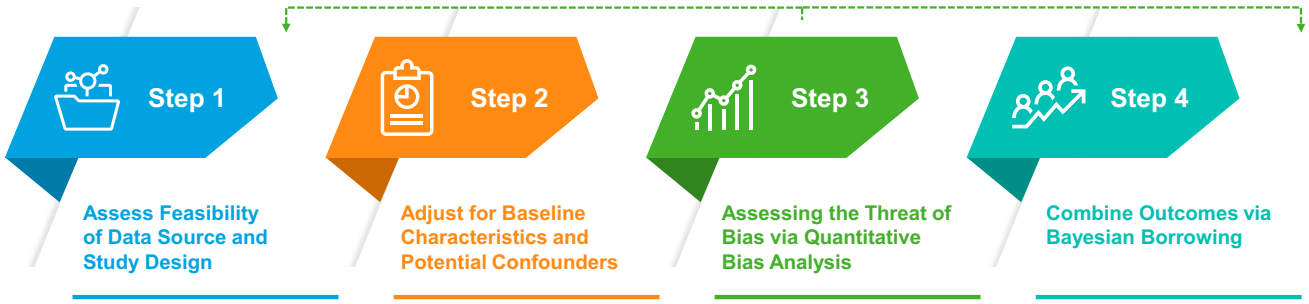

Figure 2. Four-step approach to a trial with real-world data external comparators. Step 1 includes an assessment of sources of bias and exchangeability. Step 2 adjusts for measured confounders using analytical approaches. Step 3 uses quantitative bias analysis to quantify the impact of potential bias. Step 4, applicable only to augmented randomised controlled trials, combines outcomes between the internal and external comparator groups dependent upon the similarity of the two cohorts. The *dotted line* from Step 3 indicates that this step may be re-ordered or implemented iteratively

The availability and validity (in comparison to the RCT) of variables representing each factor from the RWD source(s) should be assessed in terms of Pocock's criteria of exchangeability. Specific variables for which the populations are not exchangeable should be highlighted and later addressed in Steps 2–4 of the analysis.

For example, where key variables are missing in the potential data source, this should be noted alongside the risk of bias to the study if this variable were left out (Box 2). Assessment of the threat of bias due to the missing variable can be qualitatively categorised into high, medium or low. If further information in the literature is available to help quantify typical values of missing variables, any further quantitative impact may be assessed in Step 3. Variables with information bias in the RWD relative to the trial should be similarly be identified.

Box 2. Assessment of key variables.

For each variable, it should be assessed in the RWD whether:

The variable is collected identically to the RCT and is fully available.

The variable is partially missing or prone to misclassification or measurement error; bias analysis should be considered (Step 3).

The variable is not available or collected in a way completely insufficient to the needs of the study; consider whether the study is feasible without this variable.

Transparency and documentation in this feasibility process will likely bolster confidence in the data and the ability to identify sources of bias. Epidemiological and data source expertise are required to detect less obvious systematic differences in data collection. The outcome of this phase would be a report addressing the exchangeability criteria and highlighting differences between the two populations and data sources, which will inform quantitative bias analysis in Step 3.

6 Safety Outcomes

In studies with safety outcomes, misclassification of the adverse events due to passive reporting in RWD is of particular concern. The under-reporting and recording of non-serious adverse events in RWD is well known [35]. For example, in the evaluation of Heparesc, a treatment of ornithine-transcarbamylase deficiency in low-weight infants, the EMA assessed that a "comparison of safety data of the treated patients with historical patients is also problematic" due to "underreporting for the historical controls thus disfavouring the experimental arm" [36].

To effectively use RWD external comparators for safety outcomes, it is necessary to validate proxy variables or establish reference values for the typical rate of misclassification or measurement error. For example, one initiative by the Friends of Cancer Research is to establish the concordance of proxies for common oncology outcomes across RWD sources [37] and another to develop real-world benchmarks for metastatic pancreatic cancer [38]. In addition, several publications have called for the routine collection of RWD for all participants in RCTs, which would aid in establishing reference values [11, 26, 27].

6.1 Step 2: Adjust for Baseline Characteristics and Potential Confounders

All clinical studies strive to reduce confounding by balancing covariates between treatment groups. Common methods of balancing covariates include randomisation, use of selection criteria or analytical methods such as adjustment within the outcome regression model, aggregate constructs such as propensity scores or causal inference approaches [39]. Step 2 aims to address the differences in measured variables between the trial and external comparators, specifically selection criteria, baseline characteristics and measured confounders.

Propensity score methods are frequently used for covariate balancing in observational studies [41–44]. Propensity score matching in particular aims at replicating randomisation by matching treated patients to comparators based on their probability of being treated given patient characteristics [44]. The propensity score methodology can also be adapted to quantify and illustrate the exchangeability of the external comparator population with that of the trial population. For a supplemented SAT, there is no difference as all trial participants are treated. For an augmented RCT, one would want to compare the external comparators to the whole trial population to assess exchangeability. Therefore, one estimates the probability of a patient being in the clinical trial or the RWD, based on the measured patient characteristics. This approach is only to evaluate the exchangeability of the populations and not for adjustment of the treatment effect.

In external comparator studies, we suggest that where very low scores or little overlap in score distributions are observed (i.e. the external comparators have low probability of being included in the trial), the feasibility of using those patients as external comparators should be questioned.

Where there is overlap in the score distributions, the extent of the overlap may provide insight into the limits of generalisability of the trial.

If after adjusting for covariate imbalance there are still large differences in the outcomes between trial controls and external comparators, it indicates that the populations are only partially or poorly exchangeable owing to unmeasured factors or other sources of bias. In a supplemented SAT, we cannot compare outcomes as there are no internal controls, however, it can be assessed, modelled, and incorporated into the analysis of augmented RCTs with both internal and external controls (Step 4). Even in a supplemented SAT, however, we can quantitatively assess the threat of bias from unmeasured sources (Step 3).

6.2 Step 3: Assessing the Threat of Bias via Quantitative Bias Analysis

The practice of modelling sources of bias, using deterministic or probabilistic models, which may impact results of research has been termed QBA [45–47]. In formal QBA, one defines the model whereby the source of bias may impact the treatment effect estimate (i.e. the bias model), the minimal set of parameters governing this model (i.e. bias parameters), and feasible values for these parameters from the literature or using expert knowledge [48]. Most commonly, each potential source of bias and the impact is modelled separately [45, 49–53]. See the Electronic Supplementary Material for a simple example of the application of QBA in the context of selection bias in an external comparators study.

In other areas of epidemiology, there is increased demand for bias assessment to be integrated at both the peer review and regulatory level [54, 55], as QBA quantifies the possible magnitude, direction and uncertainty around the bias for decision makers. Regulators commonly request additional data or analysis to address bias in studies making use of external comparators (see Box 3) and QBA can be used to address this need [25, 56]. However, QBA can be undertaken as part of the analytical design or even in Step 1 alongside sample size estimates before analysis is undertaken to quantitatively assess data source feasibility.

Box 3. Examples of regulatory use of informal bias assessment

Yescarta, EMA Review:

In the 2018 EMA assessment of Yescarta, a treatment for relapsed or refractory non-Hodgkins lymphoma, a Phase II single-arm trial was described, which had an *a priori* benchmark of achieving >20% observed response rate (ORR). The historical control study, composed of control arms of two clinical trials, found that the natural history response rate was 25.7%. Assessing the potential bias from differences in the clinical characteristics of the patients, the assessors performed a re-analysis of just the "worst-case" patients. The re-analysis provided sufficient evidence that Yescarta had a superior treatment effect. This is an example of informal bias assessment by regulators.

Emtricitabine/tenofovir alafenamide, FDA Review:

In the 2015 FDA review of emtricitabine/tenofovir alafenamide, an antiretroviral therapy for HIV-1 infection, non-randomised cross-study comparisons were reported for different combinations using the therapy. The regulator listed Pocock's six criteria as a means to assess potential bias and identified that the criteria indicating that patient characteristics should be similar, was not necessarily met. Specifically, it was found that "differences in baseline characteristics were larger than would be expected in a randomized comparison but were qualitatively similar except for region of enrollment". Use of formal QBA could have evaluated quantitatively how large an impact on the outcomes that selection bias may have had. However QBAwas not performed. The FDA concluded that a fully powered RCT was necessary for emtricitabine/tenofovir and the 2015 application was not successful.

The most comprehensive form of QBA is to create a deterministic model of one or more bias mechanisms followed by stochastic modelling by varying key parameters in simulations (i.e. Monte Carlo simulations) to estimate the magnitude and uncertainty associated with bias in the treatment effect [47]. As this becomes computationally intensive as the number of potential sources of bias increases, care must be taken in choosing the largest threats of bias based on the feasibility assessment in Step 1 [47].

Another form of QBA is nullification analysis, where one estimates the 'E-value', i.e. the strength of the bias required to nullify the observed effect [57]. If very strong confounding is required to achieve nullification and subject-matter experts find confounding of that magnitude unlikely, then the risk of bias to the validity of the study would be considered minimal.

Quantitative bias analysis models may also be implemented in a Bayesian setting by formalising the range of bias estimates into Bayesian priors. Our opinion is that this

approach may be combined with Step 4, Bayesian dynamic borrowing, discussed in the next section, but this combination of methods requires further research.

6.3 Step 4: Combine Outcomes via Bayesian Dynamic Borrowing

Pocock proposed using Bayesian methods to combine the evidence from (exchangeable) historical data with the evidence from the control arm of an RCT [24]. An in-depth explanation of Bayesian methods can be found in various textbooks [58–60] and a quick review of the lexicon is provided in Box 4. These methods, referred to broadly as "Bayesian dynamic borrowing" or "Bayesian discounting functions", create priors from external comparator data, which can then be applied to the internal control data to increase the total power of the control group. As an internal control group is required, Bayesian dynamic borrowing can only be applied to augmented RCTs and not to supplemented SATs. It is of note that there is little published research applying these methods in the context of RWD. Therefore, this section is intended to be suggestive of how to evaluate methods for suitability in this context.Two general approaches for creating the prior for external comparator data are illustrated in Fig. 3. The first is the "power prior" approach [61, 62], wherein one first combines an "uninformative" prior distribution [59] with the external comparator data to create a new prior distribution (the power prior); this power prior is then combined with the likelihood of the trial data to form a posterior distribution that incorporates both the trial and external comparator information. The simplest implementation of the power prior approach includes a weighting parameter, ω, to additionally discount the external data where ω = 1 is full exchangeability and ω = 0 is non-exchangeability. A review of various implementations of the power prior can found in Viele et al. (see Fig. 3A) [31]. The second approach is the meta-analytic predictive (MAP) prior in which one or more external comparators are combined via a meta-analytic model to form a prior distribution summarising all the information [63].

Box 4. Bayesian lexicon

All Bayesian models are made of a **prior distribution**, i.e. a probability distribution representing a hypothesis or existing knowledge, a **likelihood distribution**, i.e. the current data or trial, and a **posterior distribution**, i.e. the updated hypothesis or knowledge.

A **posterior mean** of the treatment effect is analogous to the **maximum likelihood estimate** typically found in a frequentist framework. A **95% credible interval** can be constructed around the posterior mean akin to a **95% confidence interval** to quantify uncertainty.

An **uninformative prior** is any distribution which provides no information to the model. The combination of an uninformative prior and the likelihood should result in posterior estimates which have the same **type I error** rate as typical frequentist estimates.

In general, frequentist methods and drug development trials rely on ensuring a low type I error, traditionally less than 5%. Use of the above methods, when well specified can in theory result in a lower type I error in the long run. However, violations of the assumption of exchangeability as well as poor specification of priors or weighting can result in higher type I error. See Viele et al 2014 for an illustration of the effect [31].

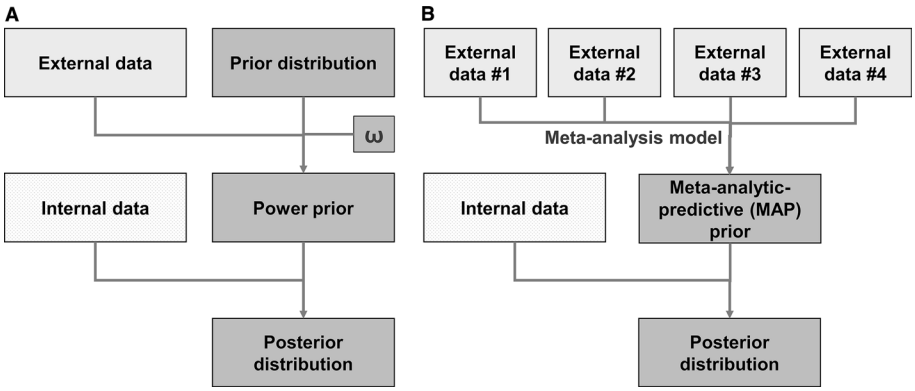

Figure 3. Basis for Bayesian dynamic borrowing. **a** The "power prior" [61, 62, 66] is constructed from an uninformative prior, the likelihood of the external comparator data and a weighting parameter ("the power parameter"; depicted as ω) used to discount the external data, accounting for the either the measured or unmeasured differences in the populations. This power prior is then applied to the likelihood of the randomised controlled trial internal control data to estimate a Bayesian posterior distribution of the outcome. **b** Bayesian hierarchical models may assume that each source of data is sampled from a larger population [31, 70]. The resulting variability between sources is modelled as a random effect whose variance is to be estimated. Many sources of data are required to accurately model the variance without robust priors

Subjectivity in the specification of parameters may be problematic. For example, in the basic power prior method scenario, ω may be specified based upon the exchangeability conclusions from Steps 1–3; however, more research would be required to establish appropriate guidance. The weight could be considered a random parameter [61, 64], but it has been shown that use of random weighting tends to overly attenuate the contribution of the external comparators [65].

Methods that more objectively account for differences in the observed data sets, i.e. borrow dynamically depending on the heterogeneity of the data, have been developed. These include the modified (normalised) power prior [66], the commensurate power prior [67, 68] and the MAP prior [63]. These methods are compared in simulation studies in van Rosmalen et al. [69] alongside the original power prior [61] and Pocock's method from 1976 [24]. The methods that account for heterogeneity between data sources, and therefore the degree of observable non-exchangeability in the data, were found by van Rosmalen et al. to provide the best trade-off in terms of power and type I error with the MAP prior showing the greatest promise.

Methods that objectively account for heterogeneity in the data are only accounting for observable non-exchangeability in the outcome and confounders (whereas covariate adjustment methods account for non-exchangeability only in the observed confounders). Important variables that are not measured or are measured differently in a manner that goes undetected may still introduce an element of non-exchangeability and therefore bias. For this reason, when using RWD, methods that allow for equal weighting of the external comparators to the internal control arm may be less desirable.

Application of the methods to only the outcome distributions or to aggregated data, not including the whole measured data set, is at a disadvantage as heterogeneity in the confounders is not incorporating into the discounting function. Use of traditional covariate adjustment methods to predict the outcomes dependent upon inclusion in the trial may be used in advance of use of aggregate dynamic borrowing methods, but further research comparing this approach to dynamic borrowing approaches using the full data is necessary.

Prior specification may also be subjective and can have a large effect on a Bayesian analysis [31], thus sensitivity analysis of the priors should be standard practice [71]. Informative priors should only be used in very special situations. However, we also suggest using QBA for a sensitivity analysis, varying the reliability of specific parameters within the Bayesian model to reflect possible bias.

7 Conclusions

In light of the recent increase in regulatory approvals of products on the basis of uncontrolled studies [23, 72], having an appropriate framework for evaluating data from SATs or augmented RCTs using RWD external comparators is of urgent importance. While there is discordance in regulatory decision making, the need for robust evidence and rigorous assessment of bias is clear. In this article, we have proposed a framework whereby trials including RWD external comparators may be designed in a more deliberate manner and evaluated for the quality of the evidence. This framework is based on study design (augmented RCT or supplemented SAT) and careful evaluation of the non-exchangeability between the RWD external comparators and trial population.

None of the methods presented in this article are novel; however, we have put them together in a framework to formalise their combined use. We have shown that regulators have cited exchangeability concerns directly and performed an informal bias analysis. There is a concern that only 'miracle' drugs can benefit from the use of external comparators. We believe the use of QBA can be used to better assess, both before and after the conduct of the trial, whether a trial would benefit from additional external comparator data. Where the line is between what is possible with external comparators and what is not will depend on the effect size of the therapy, the sample size and the quality of the data/impact of bias. Assessing this in advance can lead to better decisions on what therapies can make efficient use of this trial design.

However, there are still many challenges and gaps in this area. There have not been enough applications to regulators containing RWD external comparators to make an assessment of their response, nor has this framework been applied in any such applications. While others have called for greater use of QBA in regulatory decisions, the effect of its use in more robust applications on submission outcomes is yet to be discovered.

While the MAP method was shown to be quite promising [69], methods of dynamic borrowing for external comparator studies is a developing field. More research and simulations in the setting of RWD are required to demonstrate the efficacy of these methods. Furthermore, while we suggest the use of dynamic borrowing as the last step in the four-step process, the best implementation in RWD is still an area for further research.

Acknowledgements We acknowledge the contributions of the late Sir Alasdair Breckenridge, University of Liverpool, in advising us on regulatory approvals relying on external comparators in the EMA and directing us to other useful information from European regulators. We would also like to acknowledge the contributions of our colleague Ruben Hermans at IQVIA.

Compliance with Ethical Standards

Funding No other sources of funding were used to assist in the preparation of this article.

Conflict of interest Christen Gray, Fiona Grimson, Deborah Layton and Joseph Kim are employees of IQVIA but have no other conflicts of interest that are directly relevant to the content of this article. Stuart Pocock has no conflicts of interest that are directly relevant to the content of this article.

Data sharing Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Eichler H-G, Baird L, Barker R, et al. From adaptive licensing to adaptive pathways: delivering a flexible life-span approach to bring new drugs to patients. Clin Pharmacol Ther. 2015;97(3):234–46.
2. Liberti L, Bujar M, Breckenridge A, et al. FDA facilitated regulatory pathways: visualizing their characteristics, development, and authorization timelines. Front Pharmacol. 2017;8:161.
3. Pullman D, Wang X. Adaptive designs, informed consent, and the ethics of research. Control Clin Trials. 2001;22(3):203–10.
4. Mehta C, Gao P, Bhatt DL, Harrington RA, Skerjanec S, Ware JH. Optimizing trial design: sequential, adaptive, and enrichment strategies. Circulation. 2009;119(4):597–605.
5. Nelson NJ. Adaptive clinical trial design: has its time come? J Natl Cancer Inst. 2010;102(16):1217–8.
6. Lang T. Adaptive trial design: could we use this approach to improve clinical trials in the field of global health? Am J Trop Med Hyg. 2011;85(6):967–70.
7. Yin G, Lam CK, Shi H. Bayesian randomized clinical trials: from fixed to adaptive design. Contemp Clin Trials. 2017;59:77–86.
8. Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. Ann Oncol. 2017;28:34–433.
9. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N. Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999–2014. BMJ Open. 2016;6(6):e011666.
10. Pontes C, Fontanet JM, Vives R, et al. Evidence supporting regulatory-decision making on orphan medicinal products authorisation in Europe: methodological uncertainties. Orphanet J Rare Dis. 2018;13(1):206.
11. Najafzadeh M, Gagne JJ, Schneeweiss S. Synergies from integrating randomized controlled trials and real-world data analyses. Clin Pharmacol Ther. 2017;102(6):914–6.
12. Franzén S, Janson C, Larsson K, et al. Evaluation of the use of Swedish integrated electronic health records and register health care data as support clinical trials in severe asthma: the PACEHR study. Respir Res. 2016;17(1):152.
13. Committee for Proprietary Medicinal Products. Note for guidance on choice of control group in clinical trials. CPMP/ICH/364/96. 2001. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-10-choice-control-group-clinical-trials-step-5_en.pdf. Accessed 21 Apr 2020.
14. FDA Center for Drug Evaluation and Research. Summary review of Alecensa: application number 208434Orig1s000. 2015. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2015/208434Orig1s000Approv.pdf. Accessed 21 Apr 2020.
15. Committee for Medicinal Products for Human Use. Assessment report of Alecensa. EMA/CHMP/833519/2017. 2017. https://www.ema.europa.eu/en/documents/variation-report/alecensa-h-c-4164-ii-0001-epar-assessment-report_en.pdf. Accessed 21 Apr 2020.
16. European Medicines Agency. Assessment report of BLINCYTO. EMA/CHMP/469312/2015. 2015. https://www.ema.europa.eu/en/documents/assessment-report/blincyto-epar-public-assessment-report_en.pdf. Accessed 21 Apr 2020.
17. Committee for Medicinal Products for Human Use. Assessment report of Zalmoxis. EMA/CHMP/589978/2016. 2016. https://www.ema.europa.eu/en/documents/assessment-report/zalmoxis-epar-public-assessment-report_en.pdf. Accessed 21 Apr 2020.
18. Head of Medicines Agency and European Medicines Agency. HMA-EMA Joint Big Data Taskforce: summary report. 2019. https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf. Accessed 21 Apr 2020.
19. Coorevits P, Sundgren M, Klein GO, et al. Electronic health records: nNew opportunities for clinical research. J Intern Med. 2013;274(6):547–60.
20. Head of Medicines Agency and European Medicines Agency. Guideline on good pharmacovigilance practices (GVP). Module VIII: post-authorisation safety studies (Rev 3). EMA/813938/2011 Rev 3*. 2017. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post-authorisation-safety-studies-rev-3_en.pdf. Accessed 21 Apr 2020.
21. Human Medicines Development and Evaluation. Concept paper on extrapolation of efficacy and safety in medicine development. EMA/129698/2012. 2013. https://www.ema.europa.eu/en/documents/scientific-guideline/concept-paper-extrapolation-efficacy-safety-medicine-development_en.pdf. Accessed 21 Apr 2020.
22. Committee for Medicinal Products for Human Use. Guideline on clinical trials in small populations. CHMP/EWP/83561/2005. London: EMA; 2006. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-trials-small-populations_en.pdf. Accessed 21 Apr 2020.
23. Gottlieb S. Statement from FDA Commissioner Scott Gottlieb, M.D., on FDA's new strategic framework to advance use of real-world evidence to support development of drugs and biologics. FDA press release. December 2018. https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-new-strategic-framework-advance-use-real-world. Accessed 21 Apr 2020.
24. Pocock S. The combination of randomized and historical controls in clinical trials. J Chronic Dis. 1976;29:175–88.
25. US Department of Health and Human Services Food and Drug Administration. Statistical review and evaluation of emtricitabine/tenofovir alafenamide NDA 208215. 2015. https://www.fda.gov/media/98523/download. Accessed 21 Apr 2020.
26. Hemkens LG. How routinely collected data for randomized trials provide long-term randomized real-world evidence. JAMA Netw Open. 2018;1(8):e186014.
27. Mc Cord KA, Al-Shahi Salman R, Treweek S, et al. Routinely collected data for randomized trials: promises, barriers, and implications. Trials. 2018;19(1):29.
28. Lasch F, Weber K, Chao MM, Koch A. A plea to provide best evidence in trials under sample-size restrictions: the example of

pioglitazone to resolve leukoplakia and erythroplakia in Fanconi anemia patients. Orphanet J Rare Dis. 2017;12(102):1–6.
29. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Pract Epidemiol. 2016;183(8):758–64.
30. Lodi S, Phillips A, Lundgren J, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. Am J Epidemiol. 2019;188(8):1569–77.
31. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat. 2014;13(1):41–544.
32. Carpenter JR, Kenward MG. Multiple imputation and its application. 1st ed. New York: Wiley; 2013.
33. Kaji AH, Schriger D, Green S. Looking through the retrospectoscope: reducing bias in emergency medicine chart review studies. Ann Emerg Med. 2014;64(3):292–8.
34. Haneuse S, Bogart A, Jazic I, et al. Learning about missing data mechanisms in electronic health records-based research: a survey-based approach. Epidemiology. 2016;27(1):82–90.
35. Golder S, Loke YK, Wright K, Norman G. Reporting of adverse events in published and unpublished studies of health care interventions: a systematic review. PLoS Med. 2016;13(9):e1002127. Epidemiology. 2016;27(1):82–90.
36. European Medicines Agency. Assessment report of Heparesc. EMEA/H/C/003750/0000. 2015. https://www.ema.europa.eu/en/documents/assessment-report/heparesc-epar-public-assessment-report_en.pdf. Accessed 21 Apr 2020.
37. Stewart M, Norden AD, Dreyer N, Henk HJ. An exploratory analysis of real-world end points for assessing outcomes among immunotherapy-treated patients with advanced non-small-cell lung cancer special article abstract. JCO Clin Cancer Inform. 2019;3:1–15.
38. Philip PA, Chansky K, LeBlanc M, et al. Historical controls for metastatic pancreatic cancer: benchmarks for planning and analyzing single-arm phase II trials. Clin Cancer Res. 2014;20(16):4176–85.
39. Hernán MA, Robins JM. Causal inference: what if. Boca Raton: Chapman & Hall/CRC; 2020.
40. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399–424.
41. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–107.
42. Stuart EA. Matching methods for causal inference: aA review and a look forward. Stat Sci a Rev J Inst Math Stat. 2010;25(1):1–21.
43. King G, Nielsen R. Why propensity scores should not be used for matching. Polit Anal. 2016; 1–20.
44. Faria R, Hernandez Alava M, Manca A, Wailoo A. NICE DSU technical support document 17: the use of observational data to inform estimates of treatment effectiveness for technology appraisal: methods for comparative individual patient data. 2015. https://nicedsu.org.uk/wp-content/uploads/2016/03/TSD17-DSU-Observational-data-FINAL.pdf. Accessed 21 Apr 2020.
45. Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol. 1996;25(6):1107–16.
46. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413–9.
47. Lash TL, Fox MP, Maclehose RF, Maldonado G, Mccandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969–85.
48. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer; 2009.
49. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med. 2014;33(12):2137–55.
50. Gustafson P. Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments. London: CRC Press; 2003.
51. Goldstein H, Kounali D, Robinson A. Modelling measurement errors and category misclassifications in multilevel models. Stat Model. 2008;8(3):243–61.
52. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. Stat Med. 1994;13(12):1265–82.
53. Jurek AM, Maldonado G, Greenland S. Adjusting for outcome misclassification: the importance of accounting for case-control sampling and other forms of outcome-related selection. Ann Epidemiol. 2013;23(3):129–35.
54. Fox MP, Lash TL. On the need for quantitative bias analysis in the peer-review process. Am J Epidemiol. 2017;185(10):865–8.
55. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. Am J Public Health. 2016;106(7):1227–300.
56. Committee for Medicinal Products for Human Use. Assessment report of YESCARTA EMEA/H/C/004480/0000. 2018. https://www.ema.europa.eu/en/documents/assessment-report/yescarta-epar-public-assessment-report_en.pdf. Accessed 21 Apr 2020.
57. Vanderweele TJ, Ding P. Sensitivity analysis in observational research: introducing the e-value. Ann Intern Med. 2017;167:268–74.
58. Kruschke JK. Doing Bayesian data analysis: a tutorial with R and BUGS. New York: Academic Press; 2010.
59. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 3rd ed. London: Chapman & Hall/CRC; 2013.
60. Congdon P. Bayesian statistical modelling. 2nd ed. West Sussex: Wiley; 2006.
61. Ibrahim JG, Chen M-H. Power prior distributions for regression models. Stat Sci. 2000;15(1):46–60.
62. Duan Y, Smith EP, Ye K. Using power priors to improve the binomial test of water quality. J Agric Biol Environ Stat. 2006;11(2):151.
63. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 2014;70(4):1023–32.
64. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Anal. 2012;7(3):639–74.
65. Neelon B, O'Malley AJ. Bayesian analysis using power priors with application to pediatric quality of care. J Biom Biostat. 2010;1(1):1–9.
66. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. Stat Med. 2009;28(28):3562–6.
67. Hobbs BP, Carlin BP, Sargent DJ. Adaptive adjustment of the randomization ratio using historical control data. Clin Trials. 2013;10(3):430–40.
68. Murray TA, Hobbs BP, Lystig TC, Carlin BP. Semiparametric Bayesian commensurate survival model for post-market medical device surveillance with non-exchangeable historical data. Biometrics. 2014;70(1):185–91.
69. van Rosmalen J, Van DD, van Norden Y, Van LE, Lo B. Including historical data in the analysis of clinical trials: iIs it worth the effort? Stat Methods Med Res. 2018;27(10):3167–82.
70. Pennello G, Thompson L. Experience with reviewing Bayesian medical device trials. J Biopharm Stat. 2008;18(1):81–115.
71. Hoff PD. A first course in Bayesian statistical methods, vol. 580. New York: Springer; 2009.
72. US Food and Drug Administration. Framework for FDA's real-world evidence Program. 2018. https://www.fda.gov/media/120060/download. Accessed 21 Apr 2020.