







# Quantitative bias analysis for external control arms using real-world data in clinical trials: a primer for clinical researchers

Kristian Thorlund<sup>\*1</sup> , Stephen Duffield<sup>2</sup>, Sanjay Popat<sup>3</sup> , Sreeram Ramagopalan<sup>4</sup> ,  
Alind Gupta<sup>5</sup>, Grace Hsu<sup>5</sup> , Paul Arora<sup>6</sup>  & Vivek Subbiah<sup>7</sup> 

<sup>1</sup>Dept. Health Research Methods, Evidence, & Impact, McMaster University, ON, Canada

<sup>2</sup>National Institute for Health & Care Excellence, Manchester, UK

<sup>3</sup>Royal Marsden Hospital, Imperial College, London, UK

<sup>4</sup>London School of Economics, London, UK

<sup>5</sup>Cytel Inc., Waltham, MA, USA

<sup>6</sup>Dalla Lana School of Public Health, University of Toronto, ON, Canada

<sup>7</sup>Sarah Cannon Research Institute, TN, USA

\*Author for correspondence: [thorluk@mcmaster.ca](mailto:thorluk@mcmaster.ca)

Development of medicines in rare oncologic patient populations are growing, but well-powered randomized controlled trials are typically extremely challenging or unethical to conduct in such settings. External control arms using real-world data are increasingly used to supplement clinical trial evidence where no or little control arm data exists. The construction of an external control arm should always aim to match the population, treatment settings and outcome measurements of the corresponding treatment arm. Yet, external real-world data is typically fraught with limitations including missing data, measurement error and the potential for unmeasured confounding given a nonrandomized comparison. Quantitative bias analysis (QBA) comprises a collection of approaches for modelling the magnitude of systematic errors in data which cannot be addressed with conventional statistical adjustment. Their applications can range from simple deterministic equations to complex hierarchical models. QBA applied to external control arm represent an opportunity for evaluating the validity of the corresponding comparative efficacy estimates. We provide a brief overview of available QBA approaches and explore their application in practice. Using a motivating example of a comparison between pralsetinib single-arm trial data versus pembrolizumab alone or combined with chemotherapy real-world data for RET fusion-positive advanced non-small cell lung cancer (aNSCLC) patients (1–2% among all NSCLC), we illustrate how QBA can be applied to external control arms. We illustrate how QBA is used to ascertain robustness of results despite a large proportion of missing data on baseline ECOG performance status and suspicion of unknown confounding. The robustness of findings is illustrated by showing that no meaningful change to the comparative effect was observed across several ‘tipping-point’ scenario analyses, and by showing that suspicion of unknown confounding was ruled out by use of E-values. Full R code is also provided.


**Plain language summary:** Doctors and biomedical researchers are working hard to develop new medicines for rare types of cancer, but conducting traditional, strong clinical trials can be very difficult or even wrong in these cases. To help with this, researchers are increasingly using information from real-world data such as patient records to support their studies when there isn’t much data available for comparison. This information is used to create a ‘control group’ that should be similar to the group receiving the new treatment. Such ‘control groups’ are necessary because practical and ethical challenges of randomly assigning ineffective standard of care to patients in clinical trials.


However, using real-world data comes with challenges like missing information, mistakes in measurements, and the possibility that there are other factors influencing the results that we didn’t measure. Quantitative bias analysis (QBA) is a set of methods that helps us understand and measure the errors in the data that can’t be fixed with regular statistical methods.

In this article, we talk about different ways to use QBA and show an example comparing a new treatment for a specific type of lung cancer with information from patients who received a different treatment. We use QBA to make sure our results are solid even when we don't have all the information we need about the patients at the start of the study. We share the computer code we used, so other researchers can learn from our approach.

So, this primer helps researchers understand and deal with the challenges of using real-world data in cancer studies, making sure their findings are reliable and helpful.

**Tweetable abstract:** Navigating the complexities of cancer research in rare cases is challenging. Real-world data comparison is a game-changer, but it comes with hurdles. Enter Quantitative Bias Analysis - a tool that helps us understand and correct errors in the data.

 Our article dives into this, using lung cancer as a case study. Ensuring the reliability of our findings, even when faced with missing information.

 Check it out for insights into advancing cancer research! #CancerResearch #RealWorldData #QuantitativeBiasAnalysis

First draft submitted: 27 September 2023; Accepted for publication: 14 December 2023; Published online: 11 January 2024

**Keywords:** comparative efficacy • external control arms • quantitative bias analysis • real-world evidence

The development of medicines for rare oncologic patient populations leads to many trials being non-comparative with a single-arm design, since conducting a powered randomized controlled trial would be extremely difficult or unethical. However, because regulatory and reimbursement agencies typically require some evidence of comparative efficacy, the use of external control arms based on non-trial external data (typically real-world data) is becoming frequent [1–3]. External data, however, is fraught with limitations including missing data, information bias due to real world data structure and elements differing substantially from what is required for well-designed research studies. The potential for unmeasured confounding is substantial considering not every variable that needs to be adjusted for will be available in the dataset [4].

Quantitative bias analysis (QBA) is a broad collection of approaches for modelling the magnitude of systematic errors in data that cannot otherwise be adjusted for [5]. As such they also help quantify the uncertainties in estimates due to bias. Their applications can range from simple deterministic equations to complex hierarchical models. QBA applications are gaining traction in the realm of real-world data spanning use cases in regulatory settings to health technology assessment, as recently highlighted by the National Institute of Health and Care Excellence in their real-world evidence framework [2,6,7]. Their application could become increasingly useful with the growth in external control arm applications in clinical trials.

This article is intended as an introductory ‘nuts and bolts’ primer for clinical researchers and data analysts who are interested in learning about the value of QBA, when it is applied and when not, as well as the concepts and methods underlying its application. Throughout the article we use an illustrative example from a previously published methods application. No new findings are presented in this article. However, R-code is provided in the [Supplementary Material](#) for the provided examples. Lastly, this article summarizes educational material we have developed in providing several workshops, lectures and issues panels on the topic over the past two years.

### How does QBA apply to external controls?

External control arms using real-world data are typically constructed to match to the clinical trial population, treatment settings and outcomes [1,4]. Because real-world data is often not collected with a specific research objective in mind, data on important baseline confounding characteristics and components of outcome data can be either partially or fully missing, or their definition may be incongruent albeit correlated with those used in clinical trials. Therefore, statistical adjustments to balance patient populations by intervention groups such as propensity score weighting can be limited in their application. In this setting, QBA provides a solution for establishing the possible impact of multiple sources of confounding on estimates of interest that cannot be conventionally adjusted for by measured variables [7,8]. In particular, QBA applied to external control data allows one to obtain estimates of the impact that the suspected bias may have, not only on the control arm but also on the estimate of comparative

efficacy. In practice, QBA typically models a spectrum of informed assumptions about unknown or unmeasured confounders to evaluate the bias such confounders may have on the outcome of interest. Probably, the two most common QBA approaches for estimating the bias impact are ‘tipping point analysis’ and E-values, which are both applied using step-wise approaches. Both are described in more detail in the following sections.

### RET fusion positive advanced non-small cell lung cancer example

For illustration, we elaborate on the step-wise processes of applying QBA to external control arms using an example of real-world data from RET fusion-positive advanced non-small cell lung cancer (aNSCLC) patients to establish control outcome estimates for patients receiving pembrolizumab alone or in combination with chemotherapy [9]. In this example, single-arm data from ARROW, a multi-cohort, open-label, phase II study (NCT03037385), was available for  $n = 116$  patients after filtering for first-line treatment and availability of smoking status, disease stage, ECOG performance status (PS), and non-squamous histology. The source of real-world data was the Flatiron Health Clinico Genomics Database (CGDB) and Enhanced Data-Mart (EDM). Since RET fusion is rare (1–2% of NSCLC patients) and only  $n = 10$  eligible patient could be identified from the CGDB, external control patients were instead included from the EDM under similar eligibility criteria, but with no genomic data, using the assumption that the RET fusion oncogene was not prognostic based on published data (the validity of this assumption is not addressed with QBA in this article) [9]. In this example, both ECOG PS and other covariates were missing for a substantial number of patients. Since ECOG PS is a powerful prognostic factor in cancer, these data made an ideal candidate for applying QBA to explore potential bias impact that could arise from data missingness. Throughout this article we illustrate QBA applied to this dataset, focusing on overall survival as the primary end point.

### Constructing external control arms

The methodology behind constructing external control arms external real-world data has been described extensively in the literature [4]. Over the past decade, the analytic methods for incorporating external control data in comparative efficacy evaluations have evolved much beyond naive comparisons between the independent intervention group and the control group data. Concepts and terminology have been well-established and those widely used in this article are described in detail in the [Supplementary Material](#). In the vast majority of external control arm applications to date, the preferred method is inverse probability treatment weighting (IPTW) under which propensity scores are obtained for each patient in the external control group and patients are weighted in the final analysis according to their propensity for deviating from the data characteristics of the intervention group. As such, the intervention group and the external control arm obtain similar distributions of known prognostic factors. This is important because imbalance in prognostic factors are well known to cause bias in comparative effect estimates between treatments [10–12].

It should be noted that several other analytical approaches and modification of conventional IPTW are available for approximating balance in prognostic factors. These include (but are not limited to): nearest pairing by propensity scores; focusing on overlap in the distribution of propensity scores; general boosted models relying on machine learning techniques to estimate propensity scores such as boosted regression trees. For simplicity and relevance to common and current practice, we focus our attention and examples around IPTW applications, and note that the application of QBA will be highly similar with use of ‘alternative’ propensity scoring methods.

Propensity score weights can be constructed from multiple covariates. With the aid of a causal diagram, suspected confounders can be distinguished from mediators (i.e., variables that sit on the pathway between the independent and dependent variable) and colliders (i.e., variables affected both by the independent variable and another variable, and would cause bias if adjusted for). Confounders should be adjusted for whereas colliders and mediators (when the total effect of exposure is examined) should not be adjusted for.

Once the suspected confounding covariates have been identified, the similarity in the distribution of the covariates between the intervention (and control) groups is examined. Typically, similarity between the two groups is determined for each covariate independently by assessing the standardized mean difference (SMD) between the two groups. The SMD is preferred over p-values because of its robustness to sample size [13,14]. For covariates with more than three levels, a Mahalanobis distance-based method was used to generalize the SMD metric. The SMD for a covariate is calculated as the mean difference for the covariate between the two groups, divided by the standard deviation of the mean difference. Most commonly, balancing with IPTW (or other balancing methods) is performed collectively for all covariates with SMDs larger than thresholds like 0.1 or less conservatively 0.2

Table 1. SMDs for each covariate for the two comparisons both before and after IPTW matching.

| Covariate        | Unadjusted                         |                                       | IPTW-adjusted                      |                                       |
|------------------|------------------------------------|---------------------------------------|------------------------------------|---------------------------------------|
|                  | Pralsetinib vs pembrolizumab alone | Pralsetinib vs pembrolizumab w. chemo | Pralsetinib vs pembrolizumab alone | Pralsetinib vs pembrolizumab w. chemo |
| Age              | 0.655                              | 0.400                                 | 0.230                              | 0.015                                 |
| Sex              | 0.011                              | 0.187                                 | 0.072                              | 0.007                                 |
| Smoking history  | 1.310                              | 1.250                                 | 0.192                              | 0.017                                 |
| ECOG PS          | 0.050                              | 0.191                                 | 0.075                              | 0.037                                 |
| Time to 1st dose | 0.054                              | 0.148                                 | 0.078                              | 0.042                                 |
| Stage            | 0.304                              | 0.013                                 | 0.038                              | 0.028                                 |
| Race             | 0.612                              | 0.573                                 | 0.199                              | 0.061                                 |
| Metastasis       | 0.368                              | 0.333                                 | 0.241                              | 0.383                                 |

w. chemo: With chemotherapy.

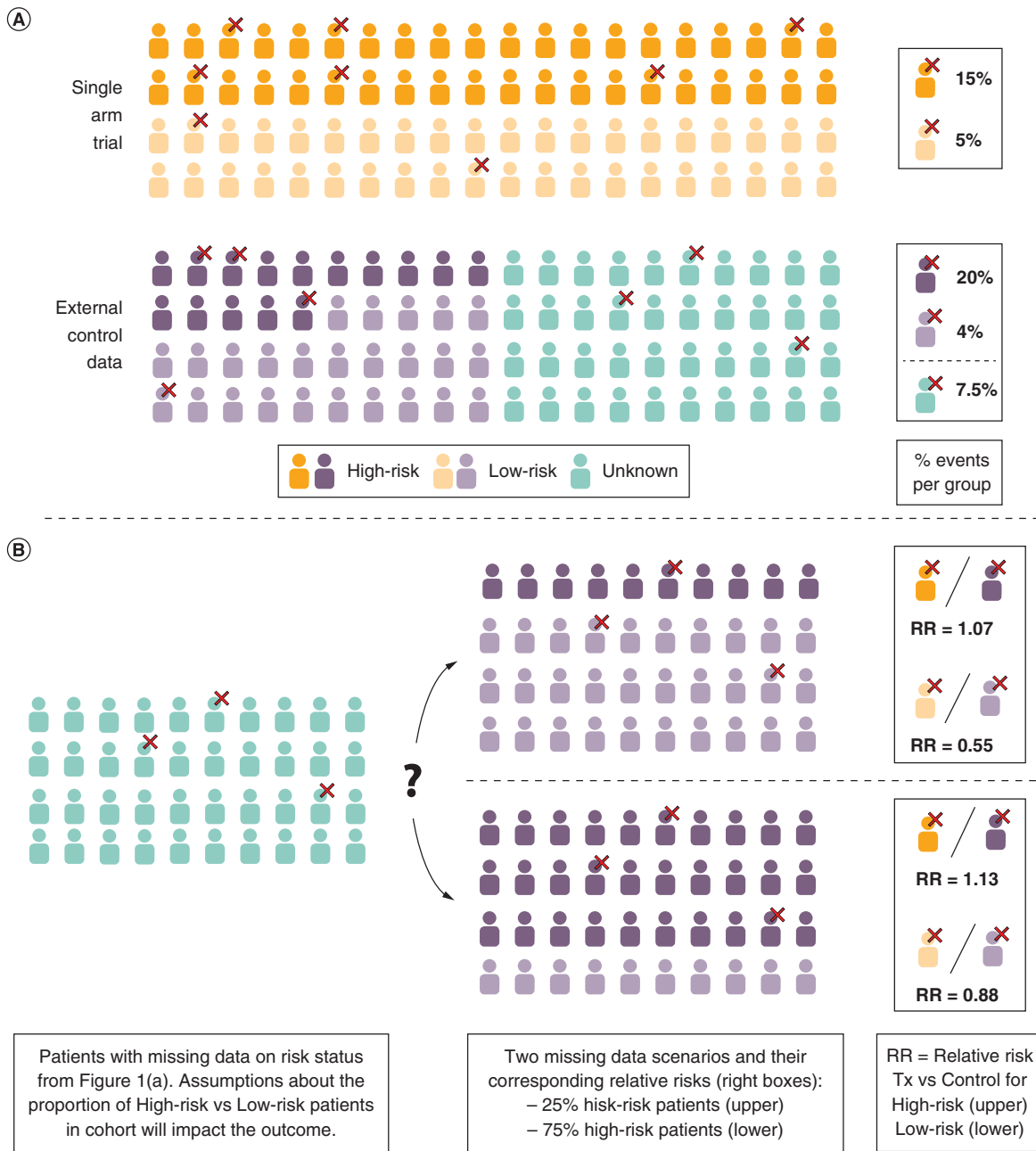
(i.e., as per Cohen's effect size index [15]) [13]. It should be noted that such thresholds are somewhat arbitrary and based on statistician authors' personal experience. SMDs are still sensitive to sample sizes (albeit less than p-values), and so, more conservative thresholds may be advised in low precision settings. It is also important to take into account whether reduction in the SMD from the non-balanced to the IPTW balanced analyses is of a meaningful magnitude. For example, a reduction from 0.44 to 0.11 represents a fourfold reduction despite not falling below the 0.1 threshold. Conversely, a reduction from 0.12 to 0.09 may not represent a meaningful reduction in covariate imbalance despite the IPTW-based SMD falling below the 0.1 threshold.

Since IPTW down weights the contribution of each patient to any percentage between 0% and 100% of the patient's unadjusted contribution, it is important to also calculate the effective sample size (ESS) after weighting [16,17]. The effective sample size is calculated as the sum of weights squared, divided by the sum of squared weights [17]. The ESS is always smaller than the actual sample size (i.e. the actual number of patients in the cohort), and the statistical precision of an IPTW adjusted cohort corresponds to that of a sample size equal to the ESS. As such, common measures of uncertainty like the 95% confidence interval always widen to the degree that the ESS is smaller than the actual sample size [17].

In our scenario of first line treatment for aNSCLC, SMDs between trial pralsetinib and real-world data pembrolizumab patients were obtained for several covariates: Age (<65 years vs >65 years), sex, smoking history, ECOG performance status (0 or 1), months from initial diagnosis to first dose, stage at initial diagnosis (I, II, III, or IV), and Race. All of these are widely recognized potential confounders in lung cancer. SMDs were obtained for comparison of covariates for both pralsetinib versus pembrolizumab alone, and for pralsetinib versus pembrolizumab with chemotherapy. Table 1 shows the SMDs for each covariate for the two comparisons both before and after IPTW balancing. Before balancing, it is clear that the SMD is consistently high for several covariates in both groups: age, smoking history, race and brain/CNS metastasis. The SMD also exceeds 0.1 for all other variables in at least one of the two comparisons. After weighting, balance was seen to be achieved. For the comparison of pralsetinib versus pembrolizumab with chemotherapy all SMDs were below 0.1 in the adjusted data. For the comparison of pralsetinib versus pembrolizumab alone, SMDs larger than 0.1 were still seen for Age, Sex and Race, albeit smaller than the unadjusted analysis. The effective sample size was ESS = 115 (compared with the original n = 683) for the pembrolizumab alone group and ESS = 217 (compared with the original n = 1270) for the pembrolizumab with chemotherapy group.

### Quantitative bias analysis approaches

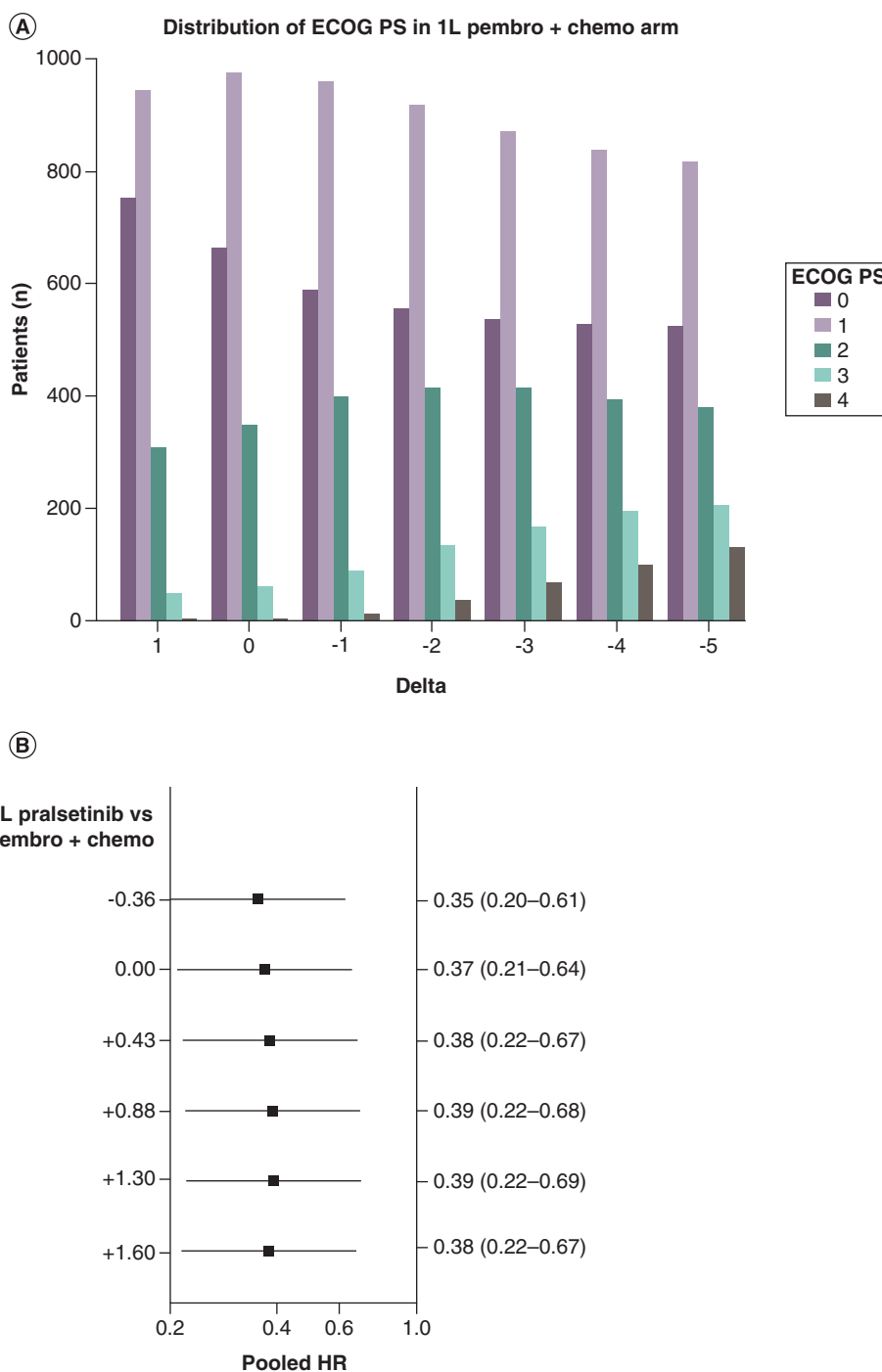
Classic epidemiological QBA guidelines by Lash *et al.* divide QBA techniques into four major categories: (1) simple bias analysis; (2) multidimensional analysis; (3) probabilistic analysis; and (4) multiple bias modelling [5]. Here (1 to 3) represents various approaches to testing one suspected confounder (1 or 3) or one at a time (2), whereas (4) represent evaluating multiple confounders under one model. Each are either applied deterministically (1 and 2), probabilistically (3), or could be applied in either way (4). In deterministic QBA, the spectrum of bias scenarios is evaluated by inputting assumed fixed values one by one for each covariate (e.g. the proportion of patients with a binary prognostic factor). By contrast, in probabilistic QBA, the spectrum of bias scenarios is evaluated by assigning probabilistic distributions around each covariate. Figure 1 illustrates a conceptual example of simple deterministic



**Figure 1. Conceptual missing data bias example.** Conceptual example of **(A)** substantial missing data on patient risk status in an external control dataset and **(B)** the changing subgroup effects (relative risks [RRs]) under two scenarios for the distribution of missing risk profiles.

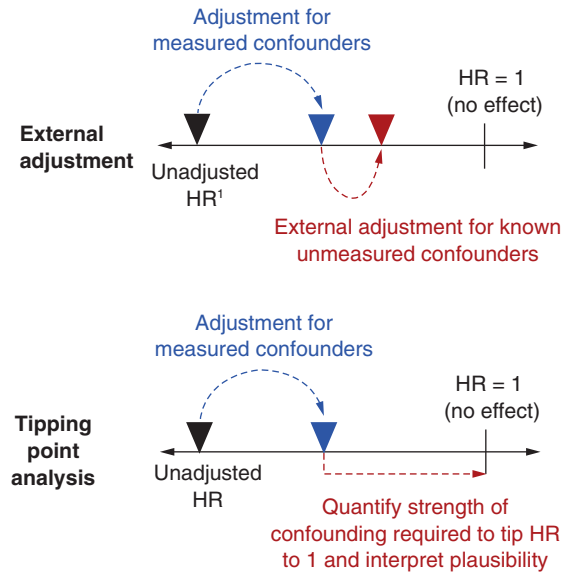
bias analysis because the analysis is focused on one unknown confounder (high risk vs low risk) and two pre-set scenarios for the external control data. Figure 2 illustrates an applied example of probabilistic QBA because the impact of multiple probability distributions for ECOG PSA on the comparative effect are evaluated. Multiple bias modelling is the most complex of the four (especially if evaluated probabilistically) as it involves evaluating multiple spectrums for multiple suspected confounders simultaneously under one statistical model.

This primer focuses predominantly on deterministic QBA, albeit with some coverage of simpler probabilistic approaches (e.g., Figure 2). Probabilistic models also require in-depth understanding of advanced statistics and Bayesian concepts. To date (December 2023), the majority of external control arm analyses have been simpler frequentist models, which are more readily subjected to deterministic QBA. Lastly, this primer also does not cover



**Figure 2. Tipping point analysis by missing ECOG data distribution.** Presents (A) the assumed distributions of ECOG values corresponding to the tested delta values; as well as (B) the HRs and 95% CIs of pralsetinib versus pembrolizumab with chemotherapy for each tested scenario. chemo: Chemotherapy; HR: Hazard ratio.

multiple bias modelling in more detail as examples of such applications are not conducive to the introductory purpose of this primer.



**Figure 3.** Illustration of single external adjustment versus tipping point analysis.  
HR: Hazard ratio.

### Tipping point analysis of missing covariate data

QBA typically employs ‘tipping point’ analysis, in which the assumed strength of confounding is increased (or decreased) up till the point where a meaningful change in the result is observed, or when the magnitude of the assumed confounding ceases to become plausible. Tipping point analysis has been recommended and used in both regulatory and HTA settings. The FDA, for example, denied Expanded Approval for Ezetimibe and Simvastatin based on tipping point analysis as part of a QBA [18]. Further, for a recent panel session at the ISPOR-US annual conference (Boston, USA; May 2023) one of the co-authors (SD) had identified over 20 NICE HTA submissions where tipping point analyses were applied to individual RCTs. Tipping point analysis is also included as a main function in over 21 software packages for QBA [19].

Tipping point analysis is useful when covariate data is either partially missing to a degree where the proportion of missing data exceeds that which is usually conducive to conventional multiple imputation (e.g., >10% or substantially larger depending on the data scenario) [20–22]. Missing data scenarios can vary substantially, and can presence or absence of biases from missing data. Since there is no clear threshold for when the proportion of missing data is important, tipping point can help evaluate whether the missingness is associated with an important bias impact in the given situation. Tipping point analysis is also highly useful for testing biases from known but unmeasured confounders. The approach is the same as for partially missing data, with the only difference being that missingness equates 100%. A common application of tipping point analysis is to increase the assumed strength of confounding until the result becomes statistically non-significant or until an effect estimate (e.g. the hazard ratio) crosses the null or some threshold for what is considered a minimally clinically meaningful effect. Figure 3 illustrates tipping point analysis contrasted to a conceptual single external adjustment to known unmeasured confounders.

Returning to our example of aNSCLC, the proportion of missingness for ECOG PS was 30% for the pembrolizumab alone group and 26% in the pembrolizumab with chemotherapy group. Although the adjusted SMDs were below 0.1 for ECOG status, the substantial proportion of missingness in both control groups raised concerns about potential selection bias. As previously mentioned, it is also well understood that ECOG status is a powerful prognostic factor. Considering the low SMDs, random missingness was identified as the most clinically plausible scenario. In missing data terms this type of missing data is referred to as ‘missing at random’ (commonly abbreviated MAR), and means that any missing data point can be predicted with good accuracy and reliability from other available covariate data. It has been well established in the missing data literature that the most robust informative method for handling data that is ‘missing at random’ is through use of multiple imputation using chained equations (MICE) [22]. As such, MICE was applied to input missing data values for both ECOG status as well as other covariates. The statistical details of MICE and imputing data in datasets with complex missing data patterns is beyond the scope of this article. For statistically minded readers, however, MICE was implemented using the `mice()` function from the `mice` R package. Documented code for this implementation can be found in [Supplementary Material](#). See [Table 2](#) for details on R packages used in the code.

Table 2. Overview of methods and metrics with suggested R packages and code for the application of quantitative bias analysis (QBA) to synthetic control arms using real world data.

| Methodology                                | Interpretation  | R package (function)                     |
|--|---|--|
| <b>Synthetic controls</b>                  |   |  |
| Propensity scores                          | Probability of treatment assignment conditional on observed baseline covariates.  | stat (glm)                               |
| Inverse probability of treatment weighting | Uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment.   | WeightIt (weightit)                      |
| Standardized mean difference               | Metric comparing means and prevalences of baseline characteristics using standardized differences, which are ratios comparing the variance of covariates between treated and untreated subjects.  | tableone (svyCreateTableOne; ExtractSmd) |
| Effective sample size                      | Metric to express how much uncertainty and error increase is due to weighting. Kish's effective sample size is a frequently-used variant.   | svyweight (eff.n)                        |
| <b>QBA – missing data</b>                  |   |  |
| Multiple Imputation                        | “Filling” in missing data in a dataset via a procedure such as multiple imputation by chained equations, which is a robust, informative method for imputing missing data in a dataset through an iterative series of predictive models.   | mice (mice)                              |
| Delta ( $\delta$ )-shift adjustment        | Under the assumption that a variable has missingness not at random, $\delta$ -based shifts for a tipping point-based bias analysis using. For $\delta$ adjustments $\delta$ is an additive term applied to the propensity score model for the variable.   | Github                                   |
| Tipping point analysis                     | An approach to manipulate scenarios to evaluate the robustness of study results by finding relevant thresholds specifying the conditions where treatment effect conclusions may hold. Involves shifting the distribution of imputed values within the RWD arm to assess whether the corresponding adjusted HRs remained significant or not. | Github                                   |
| <b>QBA – unmeasured confounding</b>        |   |  |
| E-value                                    | Represents the minimum association of a hypothetical unmeasured confounder with treatment assignment and outcome of interest on the risk ratio scale required to nullify statistically significant results.   | Github                                   |

The tipping point analysis of missing ECOG data was rooted in deviating from the assumption of ‘missing at random’. In particular, we assumed ‘missing not at random’ (commonly abbreviated MNAR), which is another common missing data terminology and describes the setting where missingness is systematic and not (fully) predictable by available covariate data. Sensitivity to deviations were specifically tested assuming a spectrum of probability distributions for the odds of observing low to high ECOG scores. A logistic regression model was employed for imputing missing ECOG scores using a delta-adjusted pattern imputation. This model incorporates a delta parameter which represents a shift parameter for the log odds of observing higher versus lower ECOG values. We used delta values of +1, 0, -1, -2, -3 and -4, which corresponds to the relative (multiplicative) shift in odds of observed ECOG values 0.79, 1.00, +0.43, 1.53, 3.67 and 4.95. Here, 1.00 indicates no change from the observed distribution among non-missing covariates, values higher than 1.00 indicate a relative increase in odds of observing ECOG scores, and values smaller than 1.00 indicate a relative decrease in odds of observing higher ECOG scores. The assumed distributions of ECOG values corresponding to the tested delta values as well as the HRs and 95% CIs of pralsetinib versus pembrolizumab with chemotherapy for each tested scenario are presented in Figure 2. The results were nearly identical for pralsetinib versus pembrolizumab alone. Across all scenarios tested, the resulting HRs and 95% CIs showed minimal sensitivity, with HRs ranging from 0.36 to 0.41 and the upper 95% CI limits ranging from 0.65 to 0.74. Thus, we could conclude that excluding patients with missing baseline ECOG values from the analysis would not have biased the comparative findings so as to nullify findings.

### E-values for unknown confounders

Another popular metric for assessing bias impact is the *E-value* [23,24]. This approach is employed to evaluate the likely impact of unmeasured confounders. The E-value is defined as ‘... the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates’ [25]. Large E-values indicate that a large strength of unmeasured confounding is required to nullify the observed treatment-outcome association (i.e. comparative treatment effect). Similarly, small values indicate that only a small magnitude of unmeasured confounding would be sufficient to nullify the observed treatment effect. However, E-values can



be complex to interpret and there is no established consensus on what constitutes a minimum threshold for an unknown confounder to be considered meaningful under QBA [26].

Proponents of E-value typically emphasize that the E-value provides a simple estimate of the joint unmeasured confounding, needed for causal inference. Its simplicity allows researchers not trained in statistics to interpret results easily (e.g., regulators and HTA agencies). No assumptions are needed about the structure of the unmeasured confounding, and it is simple to calculate from the risk ratio. The E value is therefore easy to implement as a sensitivity analysis in observational studies. Knowledge of the disease area should help identify whether any specific unmeasured confounders can plausibly cause the potential confounding. Conversely, opponents often emphasize inconsistency in interpretation. A low E value does not always mean that an unmeasured confounder could fully explain the association but possibly only to some extent, and a high E value does not always rule out unmeasured confounding. Some unmeasured confounders (e.g., those with a low prevalence) that fulfil the requirements of the E-value might not explain the observed effect. Thus the prevalence of a particular confounder should be considered carefully. Moreover, the E-value deals with confounding that inflates the magnitude of an association and cannot help evaluate confounding that masks a true association. Confounding is not the only source of bias in epidemiological studies. Therefore, the results of an E value analysis should be part of a series of sensitivity analyses addressing all threats to validity.

One way to estimate the plausibility of E-values is to use the strength of measured confounders as a proxy for unmeasured confounders, and to compare this strength to the E-value. If the E-value is larger than the strength of all measured confounders, then it may be argued that this hypothetical unmeasured confounding is not sufficient to nullify the observed treatment-outcome association [27]. Where the number of measured confounders and knowledge of their strength is limited, the E-value can either be interpreted relative to published estimates of the strength of known confounders that are not measured in the dataset, or relative to expert clinical opinion about unknown confounders. For external controls where the data is typically limited in sample size and have large proportions of missing data, E-values can be even more challenging to interpret due to imprecision, and therefore E-values should be reported for both the point estimate and the limit of the 95% confidence interval closest to null [28].

In our example of aNSCLC, using the hazard ratio for overall survival from the IPTW adjusted model, the E-value is obtained by first calculating the approximate risk ratio (RR) using the square root transformation using the following formula:  $RR \approx \frac{(1 - 0.5^{\sqrt{HR}})}{(1 - 0.5^{\sqrt{1/HR}})}$  [29]. In the pembrolizumab alone comparison, the IPTW adjusted hazard ratio was 0.29 (95%CI 0.15–0.57), and for pembrolizumab with chemotherapy the IPTW adjusted hazard ratio was 0.34 (95%CI 0.17–0.54). The E-value was estimated using the below closed form formula [25]:

$$E - value = RR + \sqrt{RR (RR - 1)}$$

For the pembrolizumab alone group, the E-value on the RR scale was 3.31, and for the pembrolizumab with chemotherapy group the E-value on the RR scale was 3.37. This means that any unknown confounder, measured on the RR scale, would need to be larger than 3.31 and 3.37, respectively, to nullify the estimated comparative effect. When no external evidence is available on the magnitude of confounding that one should expect from unknown confounders, E-values are typically contextualized based on the strength of confounding of measured confounders. In the aNSCLC example, the largest known confounders were Age and Smoking history for the pembrolizumab alone and pembrolizumab with chemotherapy groups, respectively. In particular, the observed association between Age and overall survival as the outcome was 2.10 on a risk ratio scale but little to no association was observed between age and the treatment exposure; whereas the observed association between smoking history and the treatment exposure was 7.19 on a risk ratio scale, but no association (i.e., RR = 1.00) was observed between smoking history and the outcome. Considering that the risk ratios of all considered possible confounders were generally being smaller than the obtained E-values, and further considering that any true confounder of a magnitude larger than the estimated E-values are likely to be clinically known, we can conclude that the results are fully robust against unknown confounding. R-code for the E-value analysis and evaluations are presented with documentation in the [Supplementary Material](#).

### Application to incomplete outcome data

Attempts are commonly made to construct external control arms for some clinical outcome that is not well defined or collected in the external data source. Primary outcomes that are composite (e.g., disease severity scales or collection of co-primary outcomes) in the clinical trial may not be feasible to construct from real-world data where the individual components were not necessarily collected. Similarly, definitions of disease progression in oncology often rely on established criteria (e.g., RECIST v1.1 [30]) that may not be adhered to in clinical practice. When these situations arise, QBA is best informed by extensive literature reviews on how the primary outcomes in the clinical trial correlates with the more scattered information available from external sources. Rarely will a complete mapping between these exist, and so, it is important to test plausible ranges of correlation within the QBA. This situation is akin to the aforementioned missing ECOG PS data scenario where a set of assumptions about the distributions of ECOG PS values are incorporated in the missing data model.

### Conclusion and future perspective

QBA represents an important tool for expanding the acceptability and confidence of external control arm studies, particularly when real-world data are used to constitute part, or all, of the control arm. QBA covers a rich family of analytical tools, of which we have outlined many of the simpler ones in this article. Familiarity with QBA methods such as tipping point analysis and E-values are increasing, and influential regulatory and HTA bodies are calling for their explicit use in the submission of external control arm evidence. Decision makers being called upon to assess the validity of external control arm evidence or those assigned with the design of external control arms would be well served to better understand the opportunities and challenges these methods present and how they are best implemented.

#### Executive summary

- Quantitative bias analysis (QBA) provides a methodological framework for evaluating the uncertainty around bias on clinical outcomes caused by unknown confounders, unmeasured confounding variables or partially missing covariate data.
- QBA is particularly useful for external controls arms in clinical trials, where confounding covariates from real-world data may not be recorded either fully or partially.
- QBA requires multi-disciplinary teams with in-depth knowledge of the clinical context.
- Simple QBA applications may not always suffice to capture clinical and epidemiological complexity.
- Complex applications of QBA may be time consuming and may not always add valuable knowledge about bias uncertainty.

#### Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: <https://bpl-prod.literatumonline.com/doi/10.57264/cer-2023-0147>

#### Financial disclosure

The authors have no financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

#### Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

#### Writing disclosure

No writing assistance was utilized in the production of this manuscript.

#### Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>

## References

Papers of special note have been highlighted as: ● of interest

1. Baumfeld Andre E, Reynolds R, Caubel P *et al*. Trial designs using real-world data: the changing landscape of the regulatory approval process. *Pharmacoepidemiol. Drug Saf.* 29(10), 1201–1212 (2020).
2. FDA. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products (2022). <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory>
3. Patel D, Grimson F, Mihaylova E *et al*. Use of external comparators for health technology assessment submissions based on single-arm trials. *Value Health* 24(8), 1118–1125 (2021).
4. Thorlund K, Dron L, Park JJH, Mills EJ. Synthetic and external controls in clinical trials – a primer for researchers. *Clin. Epidemiol.* 12, 457–467 (2020).
5. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int. J. Epidemiol.* 43(6), 1969–1985 (2014).
6. Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. *Am. J. Public Health* 106(7), 1227–1230 (2016).
- **This article is a great introductory read to the use and potential of quantitative bias analysis in regulatory settings.**
7. Leahy TP, Kent S, Sammon C *et al*. Unmeasured confounding in nonrandomized studies: quantitative bias analysis in health technology assessment. *J. Comp. Eff. Res.* 11(12), 851–859 (2022).
- **This article is a great introductory read to the use and potential of quantitative bias analysis in health technology assessment settings.**
8. Gray CM, Grimson F, Layton D, Pocock S, Kim J. A framework for methodological choice and evidence assessment for studies using external comparators from real-world data. *Drug Saf.* 43(7), 623–633 (2020).
9. Popat S, Liu SY, Scheuer N *et al*. Addressing challenges with real-world synthetic control arms to demonstrate the comparative effectiveness of Pralsetinib in non-small cell lung cancer. *Nat. Commun.* 13(1), 3500 (2022).
- **This real-world data study and analyses form the basis for the illustrative example in our primer and provide additional detail on analyses and data selection for the interested reader.**
10. D’Agostino RB. Estimating treatment effects using observational data. *JAMA* 297(3), 314–316 (2007).
11. Chu R, Walter SD, Guyatt G *et al*. Assessment and implication of prognostic imbalance in randomized controlled trials with a binary outcome – a simulation study. *PLOS ONE* 7(5), e36677 (2012).
12. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 319(7205), 312–315 (1999).
13. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Comm. Stat. Simul. Comput.* 38(6), 1228–1234 (2009).
14. Yang D, Dalton JE. A Unified approach to measuring the effect size between two groups using SAS. *SAS Global Forum* 335 (2012).
15. Cohen J. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Lawrence Erlbaum Associates, Publishers, NJ, USA (1998).
16. Kish L. *Survey Sampling*Wiley, NY, USA (1995).
17. Shook-Sa BE, Hudgens MG. Power and sample size for observational studies of point exposure effects. *Biometrics* 78(1), 388–398 (2022).
18. Deloughery EP, Prasad V. If the IMPROVE-IT trial was positive, as reported, why did the FDA denied expanded approval for ezetimibe and simvastatin? An Explanation of the tipping point analysis. *J. Gen. Intern. Med.* 33(8), 1213–1214 (2018).
- **This article discusses a real-world application of where tipping point analyses impacted a regulatory decision.**
19. Kawabata E, Tilling K, Groenwold RHH, Hughes RA. Quantitative bias analysis in practice: review of software for regression with unmeasured confounding. *BMC Med. Res. Methodol.* 23(1), 111 (2023).
20. Schafer JL. Multiple imputation: a primer. *Stat. Methods Med. Res.* 8(1), 3–15 (1999).
21. Bennett DA. How can I deal with missing data in my study? *Aust. NZ J. Public Health* 25(5), 464–469 (2001).
22. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*Wiley, NJ, USA (2002).
- **This book is the classic on statistical analysis with missing data and provides detailed introduction and instructions to the application of multiple imputation using chained equations (MICE).**
23. Gaster T, Eggertsen CM, Støvring H, Ehrenstein V, Petersen I. Quantifying the impact of unmeasured confounding in observational studies with the E value. *BMJ Med.* 2(1), e000366 (2023).
24. Cusson A, Infante-Rivard C. Bias factor, maximum bias and the E-value: insight and extended applications. *Int. J. Epidemiol.* 49(5), 1509–1516 (2020).
25. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* 167(4), 268–274 (2017).

- **This seminal article introduces the E-Value, provides mathematical derivations and crucial argument for the utility of E-Values**
- 26. Ioannidis JPA, Tan YJ, Blum MR. Limitations and misinterpretations of E-values for sensitivity analyses of observational studies. *Ann. Intern. Med.* 170(2), 108–111 (2019).
- 27. McGowan LDA, Greevy RA Jr. Contextualizing E-values for interpretable sensitivity to unmeasured confounding analysis. *arXiv* (2020). <https://arxiv.org/abs/2011.07030>
- 28. VanderWeele TJ, Mathur MB. Commentary: developing best-practice guidelines for the reporting of E-values. *Int. J. Epidemiol.* 49(5), 1495–1497 (2020).
- 29. VanderWeele TJ. On a square-root transformation of the odds ratio for a common outcome. *Epidemiology* 28(6), e58–e60 (2017).
- 30. Griffith SD, Tucker M, Bowser B *et al.* Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Adv. Ther.* 36(8), 2122–2136 (2019).