frontiers | Frontiers in Drug Safety and Regulation

TYPE Original Research
PUBLISHED 19 July 2024
DOI 10.3389/fdsfr.2024.1423493

OPEN ACCESS

EDITED BY
Asieh Golozar,
Northeastern University, United States

REVIEWED BY
Simone Pinheiro,
AbbVie, United States
Grammati Sarri,
Cytel, United States

*CORRESPONDENCE
Benjamin Ackerman,
✉ BAckerm3@its.jnj.com

†These authors share senior authorship

RECEIVED 25 April 2024
ACCEPTED 20 June 2024
PUBLISHED 19 July 2024

CITATION
Ackerman B, Gan RW, Meyer CS, Wang JR, Zhang Y, Hayden J, Mahoney G, Lund JL, Weberpals J, Schneeweiss S, Roose J, Siddique J, Nadeem O, Giri S, Stürmer T, Ailawadhi S, Batavia AS and Sarsour K (2024), Measurement error and bias in real-world oncology endpoints when constructing external control arms.
Front. Drug Saf. Regul. 4:1423493.
doi: 10.3389/fdsfr.2024.1423493

Measurement error and bias in real-world oncology endpoints when constructing external control arms

Benjamin Ackerman[1]*, Ryan W. Gan[1], Craig S. Meyer[1], Jocelyn R. Wang[1], Youyi Zhang[1], Jennifer Hayden[1], Grace Mahoney[1], Jennifer L. Lund[2], Janick Weberpals[3], Sebastian Schneeweiss[3], James Roose[4], Juned Siddique[5], Omar Nadeem[6], Smith Giri[7], Til Stürmer[2], Sikander Ailawadhi[8], Ashita S. Batavia[1]† and Khaled Sarsour[1]†

[1]Janssen Research and Development, LLC, A Johnson and Johnson Company, Raritan, NJ, United States, [2]Department of Epidemiology, University of North Carolina, Chapel Hill, United States, [3]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, Boston, United States, [4]Flatiron Health, New York, NY, United States, [5]Preventive Medicine and Psychiatry and Behavioral Science, Feinberg School of Medicine, Northwestern University, Chicago, United States, [6]Department of Hematology and Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, United States, [7]Division of Hematology and Oncology, Department of Medicine, University of Alabama at Birmingham, Birmingham, United States, [8]Department of Hematology, Mayo Clinic, Jacksonville, United States

Introduction: While randomized controlled trials remain the reference standard for evaluating treatment efficacy, there is an increased interest in the use of external control arms (ECA), namely in oncology, using real-world data (RWD). Challenges related to measurement of real-world oncology endpoints, like progression-free survival (PFS), are one factor limiting the use and acceptance of ECAs as comparators to trial populations. Differences in how and when disease assessments occur in the real-world may introduce measurement error and limit the comparability of real-world PFS (rwPFS) to trial progression-free survival. While measurement error is a known challenge when conducting an externally-controlled trial with real-world data, there is limited literature describing key contributing factors, particularly in the context of multiple myeloma (MM).

Methods: We distinguish between biases attributed to how endpoints are derived or ascertained (misclassification bias) and when outcomes are observed or assessed (surveillance bias). We further describe how misclassification of progression events (i.e., false positives, false negatives) and irregular assessment frequencies in multiple myeloma RWD can contribute to these biases, respectively. We conduct a simulation study to illustrate how these biases may behave, both individually and together.

Results: We observe in simulation that certain types of measurement error may have more substantial impacts on comparability between mismeasured median PFS (mPFS) and true mPFS than others. For instance, when the observed progression events are misclassified as either false positives or false negatives, mismeasured mPFS may be biased towards earlier (mPFS bias = −6.4 months) or later times (mPFS bias = 13 months), respectively. However, when events are correctly classified but assessment frequencies are irregular, mismeasured mPFS is more similar to the true mPFS (mPFS bias = 0.67 months).

Discussion: When misclassified progression events and irregular assessment times occur simultaneously, they may generate bias that is greater than the sum of their parts. Improved understanding of endpoint measurement error and how resulting biases manifest in RWD is important to the robust construction of ECAs in oncology and beyond. Simulations that quantify the impact of measurement error can help when planning for ECA studies and can contextualize results in the presence of endpoint measurement differences.

KEYWORDS

measurement error, real-world data (RWD), oncology, external control arm, misclassification bias, surveillance bias, progression-free survival, endpoints

1 Introduction

Randomized controlled trials (RCTs) remain the gold standard for the investigation of treatment efficacy (Meinert, 1996; Sibbald and Roland, 1998; Meldrum, 2000; Hariton and Locascio, 2018). However, in situations where a randomized design is not feasible due to ethical reasons, challenges in enrollment of trial participants (e.g., in cases of rare disease, or other highly specialized populations) or lack of clinical equipoise, there is an increasing interest in the construction of external control arms (ECAs) for comparison to a single-arm trial (Schmidli et al., 2020; Carrigan et al., 2022; Oksen et al., 2022; U.S. Department of Health and Human Services, Food and Drug Administration, 2023). An ECA may be identified from real-world data (RWD) sources collected outside of the trial such as electronic health records, registries, or administrative claims data. Challenges from potential biases, namely, bias due to measurement error (i.e., misclassification bias, surveillance bias), between RWD and RCTs have limited the use and acceptance of ECAs as comparators to trial populations (Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, 2021; Webster-Clark et al., 2020). Given these biases, careful consideration must be given to align populations and clinical endpoints in the trial and real-world (Seeger et al., 2020; LoCasale et al., 2021).

Oncology clinical trials commonly use a primary endpoint of progression-free survival (PFS), the earliest time from the start of treatment to a progression event or death, which is a widely accepted surrogate for overall survival (OS) (U.S. Department of Health and Human Services, Food and Drug Administration, 2018; Lin et al., 2023). Progressive disease can be measured and determined in different ways for different cancer types; for example, imaging is a common and accepted modality for assessing progressive disease in solid tumor types (Eisenhauer et al., 2009). In multiple myeloma (MM), a cancer that forms in plasma cells, disease progression is determined in accordance with the International Myeloma Working Group (IMWG) Treatment Response criteria, which is based on results of blood, urine, and bone marrow assays, as well as imaging techniques that evaluate bone lesion (Kumar et al., 2016). These key biomarkers for disease assessment are typically collected routinely in a trial setting per a protocol-defined schedule; however, in real-world settings, the collection and timing of these biomarkers and imaging assessments may vary, and their availability in RWD may be affected by data capture and normalization processes. These differences in how and when disease assessments occur in the real-world may lead to measurement error, and may limit the comparability of real-world PFS (rwPFS) to trial PFS by introducing misclassification bias and surveillance bias, respectively.

While measurement error is a known challenge when augmenting single-arm trials with RWD, there is limited literature describing factors that contribute to measurement differences (e.g., misclassified events, irregular assessment frequencies), particularly in the context of MM. Furthermore, much is still unknown regarding the potential impact of these types of measurement error in different contexts and how they interplay. The aim of this study is to investigate key sources of measurement error that contribute to bias when estimating rwPFS and illustrate how they may impact the comparability with trial PFS using a simulation example. The sections of this paper are as follows: First, we define the measurement error types of interest and provide framing for how their related biases may manifest in RWD. We then conduct a simple simulation study to illustrate how these biases may behave, both independently as well as together. We conclude by highlighting considerations on how measurement error may impact the estimation of rwPFS and discuss the importance in further quantifying bias due to these errors in practice when comparing real-world and clinical trial endpoints.

2 Methods and materials

2.1 Defining types of bias due to endpoint measurement error

In this section, we disaggregate bias of rwPFS endpoints due to measurement error into misclassification bias and surveillance bias. The former describes a bias attributed to *how* the endpoint is *derived* or *ascertained*, such that the true disease status may not be observed. The latter describes a bias attributed to *when* outcomes are *observed* or *assessed*, namely, at a different (and irregular) interval than a trial. In the context of ECAs, it is important to note that these biases are defined in *relation to the trial* population as the source of “truth.” In other words, here, we refer to biases that are attributed to differences between the RWD and the trial approaches to disease evaluation, which present when using RWD *in lieu* of a randomized trial’s control arm as a comparator. We now define these biases and discuss attributes of RWD that contribute to them and highlight how they manifest in the context of MM.

2.1.1 Misclassification bias

At each disease assessment time point, a patient's progressive disease status can be misclassified in one of two ways: *false negatives* are when patients experience progression events, but the events are not captured or observed, whereas *false positives* are when patients do not experience progression events but are falsely classified as having progressed at a certain time. These misclassification errors can impact the observed time to first progression, which in turn affects how the PFS endpoint is constructed. More specifically, false negative events may lead to longer observed PFS times, while false positive events may lead to shorter observed PFS times. It is important to note that false negatives are only possible among patients who truly progress, and thus, the impact of such errors on bias in the PFS endpoint are also dependent on the disease setting and true progression event rate. For example, in disease settings where true progression event rates within a typical duration of follow-up are low (e.g., newly diagnosed multiple myeloma, or NDMM), fewer false negative errors are possible, and therefore false positive errors are more likely to drive overall bias when estimating median PFS.

Similarly, misclassification of progression events may not always introduce bias in the PFS endpoint, and may depend on the amount of time bias attributed to the error types (Bakoyannis and Yiannoutsos, 2015; Edwards et al., 2023). For example, if a progression event is not captured (i.e., there is a false negative), but the patient has a death event that occurs a few weeks later, then the observed "mismeasured" PFS may only be biased by a few weeks. Alternatively, if a progression event is falsely detected (i.e., there is a false positive) many months before a patient truly progresses, then their PFS time may be more substantially biased.

In MM RWD, there may be high rates of missingness among biomarkers required to derive progression according to full IMWG criteria due to real-world care patterns (e.g., the urine protein electrophoresis test, UPEP, requires a patient to collect urine over 24-h), which limits real-world data use (Foster et al., 2018). Furthermore, data missingness in RWD may reflect an absence of test collection, or it may reflect tests that are collected but not observed or captured in the data source (Sondhi et al., 2023; Vader et al., 2023), and thus the full IMWG criteria for deriving progression may not be feasible to implement. Flexible alternate algorithms for deriving endpoints may be used instead; these alternative algorithms are based on IMWG criteria but are designed to be more accommodating of real-world lab collection practices or missingness rates (Foster et al., 2019). Application of these alternative algorithms may lead to misclassification of progression events relative to the full IMWG criteria as they would be applied in a clinical trial setting. While it may be possible to make minor improvements or alterations to how real-world progression is derived in MM, an "error-free" flexible algorithm may not be achievable considering differences in the underlying data availability and completeness as well as clinical practice.

2.1.2 Surveillance bias

In a clinical trial setting, patients are assessed according to a protocol-specified schedule (e.g., on a bi-monthly or monthly frequency). While it is possible for patients to have a progression event or clinical worsening in between scheduled assessments, progression events are typically detected when a patient returns for their subsequent visit. This may lead to a delay between when a progression event truly occurred and when it was observed, otherwise referred to as surveillance bias (Panageas et al., 2007). Such delays in event detection can depend on the length of the assessment intervals (i.e., if patients are assessed more frequently, the time between event occurrence and event detection may be shorter) (Kapetanakis et al., 2019; Adamson et al., 2022; Zhu and Tang, 2022).

In a randomized controlled trial, patients in both arms follow the same assessment schedule, and therefore any event detection delays may be assumed to be similar across arms. Therefore, such event detection delays may not impact treatment effect estimates. However, in the context of ECAs, patients in the external comparator may be assessed on a *different* frequency than the internal arm. Such differences in assessment schedules may lead to biased estimates when comparing the two arms.

In contrast to trials, patients in the real-world setting are not always assessed according to a strict schedule. It is possible that, on average, patient visits are distributed with some degree of consistency, albeit likely with much higher variability than in a trial. Irregular assessment frequencies may therefore be observed in RWD, and this could be for several reasons: 1) RWD patient populations are often quite heterogeneous, so there may be variations in how often patients come in for visits based on site or clinical practice, geographic proximity, or socio-economic factors, and 2) patient visits may be driven by symptoms, management of co-morbid conditions, convenience or other factors, and clinicians may recommend that patients schedule their subsequent appointments sooner or later accordingly. When conducting an externally-controlled trial with RWD, such differences in assessment timing may contribute to biased endpoint comparisons.

3 Simulation study

We now describe a simulation study to illustrate the association between misclassification of progression events, irregular assessment frequencies, and biases due to these errors in the PFS endpoint. Let N denote the total number of patients in our external comparator. For all N patients, we start by simulating true times to death (for OS), PFS and end of follow-up (FUP) using independent exponential distributions where the rates are defined by the desired median times of mOS, mPFS, and mFUP, respectively:

$$\begin{array}{ll} T_{OS} \sim Exp(\lambda_{OS}), & \lambda_{OS} = \log(2)/\mathrm{mOS} \\ T_{PFS} \sim Exp(\lambda_{PFS}), & \lambda_{PFS} = \log(2)/\mathrm{mPFS} \\ T_{FUP} \sim Exp(\lambda_{FUP}), & \lambda_{FUP} = \log(2)/\mathrm{mFUP} \end{array}$$

To simulate patients' true time from treatment initiation to first progression, T_{prog}, we compare their simulated PFS time with their simulated OS time and derive it as follows:

$$T_{prog} = \begin{cases} T_{PFS} & \text{if} \quad T_{PFS} < T_{OS} \\ NA & \text{if} \quad \text{otherwise} \end{cases}$$

Since PFS is a composite of time to death and time to first progression, this allows us to determine if the simulated PFS time is attributed to a progression event or a death event; if a patient's PFS time is earlier than their OS time, then we can infer that a progression event occurred. Otherwise, we can infer that the PFS time is attributed to a death event.

Once underlying true progression event times are simulated, the events are mapped to a fixed trial-like disease assessment schedule, such that progression events are only observed when patients are simulated to be evaluated by clinicians per protocol:

$$\bar{T}_{prog} = \left\lceil \frac{T_{prog}}{d} \right\rceil \times d$$

where d is the per-protocol time between assessments and $\lceil\ \rceil$ is the "ceiling" function that rounds the contents up to the nearest whole number. For example, if a patient's T_{prog} is simulated to occur on day 53, and they are assessed every 28 days per trial protocol (e.g., $d = 28$), then their true progression event will be observed on day $\lceil \frac{53}{28} \rceil \times 28 = 56$, since $\lceil \frac{53}{28} \rceil = \lceil 1.89 \rceil = 2$.

Patients' true PFS times are then constructed as $\min(\bar{T}_{prog}, T_{OS}, T_{FUP})$ and event indicators are determined by this time. Next, we describe how we introduce misclassification of progression events and irregular assessment times via simulation.

3.1 Simulating misclassification errors and mismeasured PFS times

To simulate misclassification errors, we assume that a flexible alternative IMWG algorithm to derive progression events with known sensitivity and specificity has been applied. We define sensitivity and specificity based on 1) whether patients' true PFS time is equal to their mismeasured PFS time (i.e., PFS constructed using progression real-world derived progression events) and 2) whether patients' true PFS time is determined by a progression event (versus death or censoring) as follows:

Sensitivity = P (true PFS = mismeasured PFS | true PFS = time to first progression)

Specificity = P (true PFS = mismeasured PFS | true PFS = time to death or censoring)

Let $N_{\text{pfs-prog}}$ denote the number of patients for whom PFS is defined by a progression event. Recall that only patients who truly have PFS defined by a progression event can be classified as a false negative. To simulate *false negatives*, we simulate $N_{\text{pfs-prog}}$ Bernoulli events with probability $p_{fn} = (1 - \text{sensitivity})$. To simulate *false positives*, we simulate $N - N_{\text{pfs-prog}}$ Bernoulli events with probability $p_{fp} = (1 - \text{specificity})$.

Next, we generate the mismeasured time to progression based on the misclassification type.

For *false negative patients*, we simulate the time as:

$$T^{mis}_{prog} = T_{prog} + T_{\text{fn bias}}$$

where $T_{\text{fn bias}} \sim Exp(\lambda_{PFS})$. In other words, for each false negative patient, we add random exponentially distributed time to their time to progression.

For *false positive patients*, we simulate the time as:

$$T^{mis}_{prog} = T_{PFS} - T_{\text{fp bias}}$$

where $T_{\text{fp bias}} \sim U[0, T_{PFS}]$. In other words, for each false positive patient, we generate a progression event that falsely happened *any time* between treatment initiation and their true PFS time.

When simulating the impact of misclassification bias only (i.e., no irregular assessment frequency), mismeasured time to progression is mapped to the trial protocol assessment schedule as $\overline{T^{mis}_{prog}} = \lceil \frac{T^{mis}_{prog}}{d} \rceil \times d$. This is then used to construct the mismeasured PFS endpoint as $\min(\overline{T^{mis}_{prog}}, T_{OS}, T_{FUP})$ and mismeasured event indicators are determined accordingly.

3.2 Simulating irregular assessment schedules and observed PFS times

As described above, we assume that disease assessments for progression in a trial follow a strict disease assessment schedule of every d days per protocol. In RWD, on the other hand, we assume that a patient is assessed on an *irregular* schedule, where the mode of time between visits is d days, but with greater variability than the trial. For example, RWD patients may be assessed *roughly* every 28 days, but may, on occasion, have visits that are more (or less) spread out (Foster et al., 2018; Roose et al., 2022). To simulate irregular times between assessments, assuming a trial-like mode with greater variability, we use a mixture of distributions that contains an identifiable mode, but with variability that may be characterized by another distribution. Here, we will use a log-normal mixture distribution:

$$g(x; \mu_1, \sigma_1, \mu_2, \sigma_2, p) = (1 - p) f(x; \mu_1, \sigma_1) + p f(x; \mu_2, \sigma_2)$$

Where g is the mixture function, f is the lognormal function, p = probability of assessment being off-cycle (i.e. deviating from the mode of d days), μ_1 and σ_1 represent the log mean day and standard deviation of the "on-assessment" day and μ_2 and σ_2 represent the log mean and standard deviation of "off-cycle" days.

For each simulated patient, we simulate a vector of assessment times according to this mixture distribution. Then, for patients who have been simulated to have a progression event, we shift their simulated true event time to equal the first irregular assessment occurring after the event. For example, if a patient's $T_{prog} = 140$ days, and we simulate irregular assessment times for them at days 27, 145, 171, 184 and 217, then their mismeasured progression event time would be shifted to day 145.

3.3 Simulation scenarios

In order to illustrate the potential impacts of and interplay between these measurement error biases, we consider a data generating model using parameters defined in Table 1, based on the control arm of a historical trial conducted among patients with NDMM that received lenalidomide and dexamethasone (Facon et al., 2021). Table 2 describes the scenarios of interest, varying frequencies of false positive (1—specificity) and false negative (1—sensitivity) errors, as well as the type of assessment frequency that patients follow. Parameter values were selected based on prior feasibility analyses and clinical perspectives regarding the performance of flexible alternative IMWG

TABLE 1 Parameters used to simulate "true" NDMM population.

Parameter	Description	Parameter value
N	Sample size	365
mOS	Median overall survival time (months)	66.4
mPFS	Median progression-free survival time (months)	34.2
mFUP	Median follow-up time (months)	56.2
end_of_study	End of study period (months)	78.6

TABLE 2 Parameters used to define mismeasurement of progression events and assessment time.

Scenario	Sensitivity	Specificity	Assessment frequency
1	1.0	1.0	Irregular, RWD-like
2	0.5	1.0	Per trial protocol, every 28 days
3	1.0	0.8	Per trial protocol, every 28 days
4	0.5	0.8	Per trial protocol, every 28 days
5	0.5	0.8	Irregular, RWD-like

TABLE 3 Simulation results varying misclassification rates and assessment frequencies.

Scenario	Sensitivity	Specificity	Assessment frequency	mPFS bias (95% CI), in months	% bias (mPFS bias/ true mPFS) (95% CI)	False discovery rate (%)
1	1.0	1.0	Irregular, RWD-like	0.67 (−7.4, 8.7)	2.0% (−21.6%, 25.4%)	5.3
2	0.5	1.0	Per trial protocol, every 28 days	13 (3.7, 22)	38% (10.8%, 64.3%)	88.2
3	1.0	0.8	Per trial protocol, every 28 days	−6.4 (−14, 0.93)	−18.7% (−41%, 2.7%)	56.8
4	0.5	0.8	Per trial protocol, every 28 days	4.8 (−3.7, 13)	14% (−10.8%, 38%)	15.9
5	0.5	0.8	Irregular, RWD-like	5.9 (−2.3, 15)	17.3% (−6.7%, 43.9%)	27.6

algorithms, as well as the frequency of patient assessments, in RWD. By studying simulation results under perfect sensitivity and specificity (Scenario 1), we can quantify biases attributed to differences in assessment frequency. By studying simulation results under trial-like assessment frequencies (Scenarios 2–4), we can quantify biases attributed to misclassification errors alone. Simulating both together (Scenario 5) will demonstrate how these biases may manifest jointly. Note that when we simulate both biases, we begin by first introducing misclassification, followed by irregular frequency.

For each simulation iteration, we generate two samples–one with the outcome measured "correctly" and the other "mismeasured" with error. We define bias as the difference in median mismeasured (i.e., "real-world") PFS and underlying true (i.e., "trial") PFS, obtained via Kaplan-Meier (KM) estimation. Positive bias denotes mismeasured PFS is longer, on average, than true PFS, while negative bias denotes mismeasured PFS is shorter than true PFS. We run 1,000 iterations of each simulation scenario and report confidence intervals as the 2.5th and 97.5th quantiles of the bias distributions. We also estimate the "False Discovery Rate" as the proportion of simulation iterations for which the true and mismeasured PFS KM curves are statistically different (defined by a p-value <0.05 via the log-rank test).

4 Results

Simulation results are presented in Table 3, and the key findings are summarized below.

4.1 Surveillance bias only (no misclassification)

Simulation scenario 1: First, let's consider the scenario where progression events are detected without any error (sensitivity and specificity = 1), but they are assessed by a clinician on an irregular frequency that is more variable than a trial protocol. Introducing

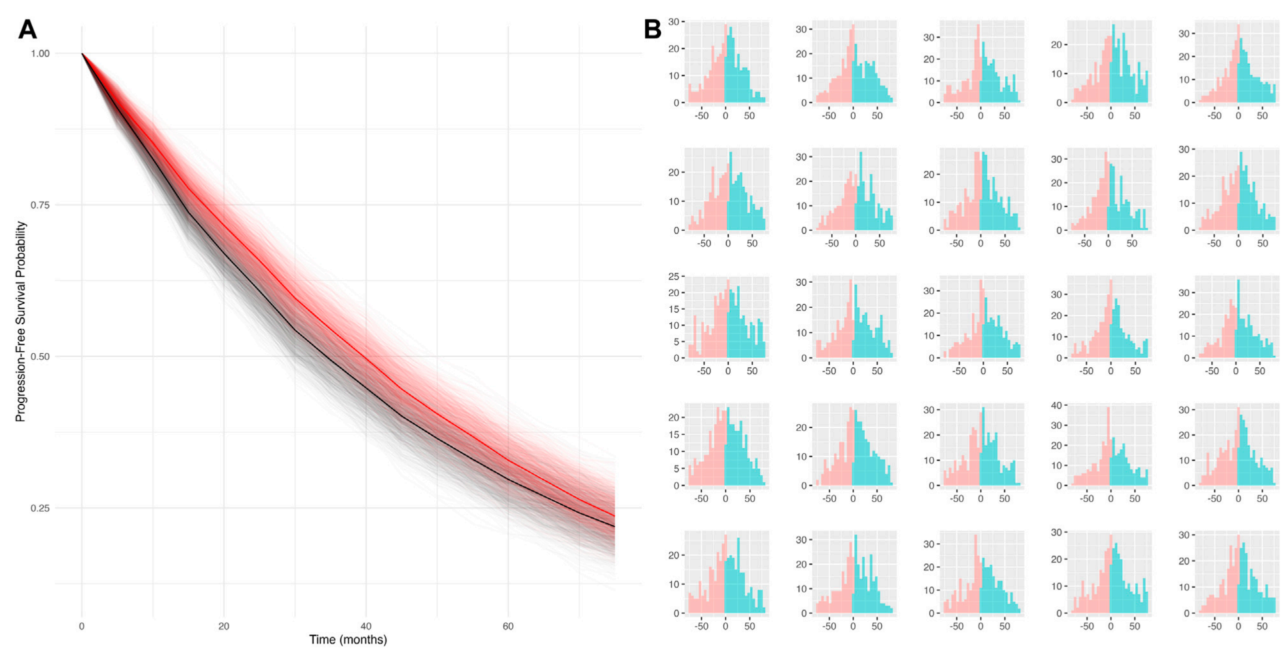

FIGURE 1
(A) Simulated KM curves comparing mismeasured (red) and true (black) PFS curves. Bold lines represent the average across all simulation runs. **(B)** Person-bias in months (true PFS−mismeasured PFS) for 25 randomly sampled simulated datasets. Bars in blue denote bias attributed to false negatives, bars in red denote bias attributed to false positives.

surveillance bias in the form of differing assessment frequencies biases the observed mPFS towards longer times (mPFS bias = 0.67 months, 95% CI: −7.4 to 8.7 months), albeit this does not translate to statistical differences between the true and mismeasured PFS curves (False Discovery Rate = 5.3%).

4.2 Misclassification bias only (no surveillance bias)

Next, let's consider scenarios where patients are assessed per a trial assessment schedule, but patients' progression events are misclassified as either false positives or false negatives.

Simulation scenario 2: When sensitivity is 50% and specificity is 100%, mismeasured mPFS is substantially biased towards longer times than the true PFS (mPFS bias = 13 months, 95% CI: 3.7 to 22 months), and there is a statistical difference between the true and mismeasured PFS curves almost 90% of the time. In other words, if a flexible alternate algorithm to derive progression in the real-world misses 50% of patients who truly progress, but does not introduce any false progression events, then such an algorithm can yield real-world mPFS that appears much longer than the truth.

Simulation scenario 3: When sensitivity is 100% and specificity is 80%, mismeasured mPFS is biased towards *shorter* times than the true PFS (mPFS bias = −6.4 months, 95% CI: −14 to 0.93 months), and there is a statistical difference between the true and mismeasured PFS curves ~57% of the time. This represents a scenario where a real-world approach to derive progression captures all true progressors but overclassifies progression for those who do not truly progress.

Simulation scenario 4: When both false positives *and* false negatives are simulated together (sensitivity = 50%, specificity = 80%), the bias is smaller (mPFS bias = 4.8 months, 95% CI: −3.7 to 13 months), and the false discovery rate is reduced (statistical differences between true and mismeasured PFS curves detected ~16% of the time). Furthermore, the PFS Kaplan-Meier curves for the true and mismeasured outcomes appear to overlap in this scenario, suggesting that the two endpoint versions may be more comparable (See Figure 1A). However, upon further inspection, these errors are each yielding their own substantial biases that appear to oppose one another (See Figure 1B).

4.3 Misclassification and surveillance bias

Simulation scenario 5: Lastly, let's consider the scenario where misclassification and surveillance biases are both present. Introducing irregular assessment frequency, on top of false positives and negatives of progression events, further biases mPFS to later times (mPFS bias = 5.9 months, 95% CI: −2.3 to 15 months), and the probability of detecting a statistical difference between true and mismeasured PFS curves increases from 15.9% to 27.6%. While irregular frequency alone had limited impact on the bias, the results in this scenario highlight that the joint contribution of misclassification and surveillance biases (the most realistic real-world scenario) may be more substantial.

5 Discussion

When developing an ECA to contextualize the findings of a single-arm trial, it is important to consider how measurement differences can affect endpoint comparability and the potential for bias. In this research, we have defined sources of measurement error that may add bias and limit comparability of real-world and trial endpoints in MM and we have highlighted several key contributors to measurement differences in PFS. Through simulation, we have illustrated how differences between real-world mPFS and trial mPFS may be attributed to misclassification bias as well as surveillance bias in this disease population.

Even when measurement error is present among individual patients, simulations suggest that such bias may not always result in observable or statistical differences between true (i.e., trial) and mismeasured (i.e., real-world) mPFS estimates. For example, simulations showed that when progression events are classified correctly, but patients have irregular assessment times (Scenario 1), a small amount of bias is observed while the false discovery rate is low in this analysis. Additionally, when patients are assessed on a regular frequency, but both false positives and false negatives are present (Scenario 4), these errors contribute large, yet *opposing*, amounts of bias that can cancel out. Under differing rates of sensitivity and specificity, it is even plausible that these errors could cancel out completely. This further raises an important cautionary point, that even if findings appear similar or unbiased between trial and RWD cohorts, biases due to measurement error may still be present at the individual level, and it is important to quantify them to contextualize such findings. However, while simulations suggest each type of bias may not substantially hinder comparability alone, it is unlikely that these phenomena exist in isolation in the real-world. When misclassified progression events and irregular assessment times occur simultaneously (Scenario 5), we have demonstrated that they can generate bias that is greater than the sum of their parts.

Note that these measurement differences may have varied effects in other disease settings with different event rates and may also depend on sample size and prognostic factors. While we assume in this illustrative simulation that the biases due to measurement error are not differential with respect to any baseline covariates, it is important for future work to study the identification and impact of important prognostic characteristics of measurement error. Furthermore, recall that we have defined these biases in RWD mPFS *in relation to* the "true" mPFS that would be observed in the control arm of a randomized trial. Future work should also examine how these biases impact treatment effect estimation (i.e., hazard ratio comparing PFS in a single-arm trial to PFS in an RWD comparator) under various effect sizes and trial outcomes.

In application, it can be challenging to quantify the amount of bias in PFS due to measurement error and definitively assess how much bias is attributed to each potential cause. In the context of MM, where real-world endpoints may be derived according to flexible algorithms using a subset of IMWG biomarkers typically observed in RWD, little is known about how such flexible algorithms perform relative to trial standards. Furthermore, different algorithms may have different false positive and negative rates, thereby yielding different amounts of measurement error and bias when compared to a trial. In this illustrative parametric simulation, we assume the performance of such flexible algorithms is already known (or estimated). Future studies should consider simulation designs that enable more direct performance evaluation of these algorithms relative to trial endpoints. Such studies would play an important role in understanding how measurement error manifests not just in theory, but in practice.

This paper highlights through simulations how measurement error and related biases may manifest and impact PFS, a time-to-event composite endpoint used in MM and oncology. Measurement error may present differently with other types of endpoints. For example, with endpoints based on binary outcomes, like overall response rate, surveillance bias may be less concerning, as it matters more *if* the outcome occurred rather than when in the study it occurred. Approaches to correct for outcome measurement error may differ for binary, continuous, or time-to-event outcomes and warrant further research (Carroll et al., 2006; Edwards et al., 2013; Kapetanakis et al., 2019; Innes et al., 2021; Zhu and Tang, 2022).

Lastly, while we have illustrated sources of measurement error in a NDMM population, mismeasured outcomes in other contexts and diseases are also common. The biases we have highlighted in this work are present and relevant across a wide range of therapeutic areas, both within and beyond oncology. Other relevant endpoints susceptible to measurement error bias may be, for example, based on imaging (Hong et al., 2012; Huang Bartlett et al., 2020; Ton et al., 2022; Mhatre et al., 2023) (i.e., in solid tumors) or patient-reported outcomes (Keogh et al., 2016). Simulation studies that quantify bias due to measurement error can be helpful tools when planning for ECA studies and can be used where possible for contextualizing study results in the presence of endpoint measurement differences. Improved understanding of the interplay between these biases in other diseases contexts may inform future approaches for mitigating measurement error biases and constructing more robust ECAs.

Data availability statement

Simulation code to generate data and conduct simulation study may be made available upon request. Inquiries can be directed to the corresponding author.

Author contributions

BA: Conceptualization, Data curation, Formal Analysis, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. RG: Methodology, Validation, Writing–original draft, Writing–review and editing. CM: Methodology, Writing–original draft, Writing–review and editing. JRW: Methodology, Validation, Writing–review and editing. YZ: Methodology, Writing–review and editing. JH: Funding acquisition, Supervision, Writing–review and editing. GM: Project administration, Writing–review and editing. JL: Methodology, Writing–review and editing. JW: Methodology, Writing–review and editing. SS: Methodology, Writing–review and editing. JR: Writing–review and editing. JS: Methodology, Writing–review and editing. ON: Writing–review and editing. SG: Writing–review and editing. TS: Writing–review and editing. SA: Writing–review and editing. AB: Funding acquisition, Supervision, Writing–review and editing. KS: Funding acquisition, Supervision, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project is supported by the Food and Drug Administration (FDA) of the United States Department of Health and Human Services (HHS) as part of a financial assistance award [FAIN] totaling $468,058 (32 percent) funded by FDA/HHS and $1,001,454 (68 percent) funded by non-government source(s). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Acknowledgments

The authors acknowledge the following consortium whose thought leadership and input aided author's efforts: Noopur Raje (Harvard), Charles Poole (UNC Chapel Hill), Issa Dahabreh (Harvard), Hemanth Kanakamedala (Janssen), Sarah Rudolph (Janssen), Joel Greshock (Janssen), John Phillips (Janssen), Yiyang Yue (Janssen), Laura Hester (Janssen), Kelly Johnson-Reid (Janssen), Jason Brayer (Janssen), and Robin Carson (Janssen).

Conflict of interest

BA, RG, CM, JW, YZ, JH, AB, and KS are employed by Janssen Research & Development, LLC and own stock in Johnson & Johnson. GM is employed by Janssen Research and Development, LLC. JR is an employee of Flatiron Health, an independent subsidiary of the Roche group, and owns stock in Roche. SS (ORCID# 0000-0003-2575-467X) is participating in investigator-initiated grants to the Brigham and Women's Hospital from Boehringer Ingelheim, Takeda, and UCB unrelated to the topic of this study. He owns equity in Aetion Inc., a software manufacturer. He is an advisor to Temedica GmbH, a patient-oriented data generation company. His interests were declared, reviewed, and approved by the Brigham and Women's Hospital in accordance with their institutional compliance policies. ON is on Advisory Boards for Bristol Myers Squibb, Johnson & Johnson, Sanofi, GPCR Therapeutics, and Takeda, receives Honorarium from Pfizer, and research funding from Johnson & Johnson and Takeda. TS receives investigator-initiated research funding and support as Principal Investigator (R01AG056479) from the National Institute on Aging (NIA), and as Co-Investigator (R01CA277756) from the National Cancer Institute, National Institutes of Health (NIH). He also receives salary support as Director of Comparative Effectiveness Research (CER), NC TraCS Institute, UNC Clinical and Translational Science Award (UM1TR004406), co-Director of the Human Studies Consultation Core, NC Diabetes Research Center (P30DK124723), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the Center for Pharmacoepidemiology (current members: GlaxoSmithKline, UCB BioSciences, Takeda, AbbVie, Boehringer Ingelheim, Astellas, and Sarepta), and from a generous contribution from Dr. Nancy A. Dreyer to the Department of Epidemiology, University of North Carolina at Chapel Hill. TS does not accept personal compensation of any kind from any pharmaceutical company. He owns stock in Novartis, Roche, and Novo Nordisk; Sikander Ailawadhi has worked as consultant for GSK, Sanofi, BMS, Takeda, Beigene, Pharmacyclics, Amgen, Janssen, Regeneron, and received research funding from GSK, BMS, Pharmacyclics, Amgen, Janssen, Cellectar, Xencor, Abbvie, and Ascentage.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adamson, B. J. S., Ma, X., Griffith, S. D., Sweeney, E. M., Sarkar, S., and Bourla, A. B. (2022). Differential frequency in imaging-based outcome measurement: bias in real-world oncology comparative-effectiveness studies. *Pharmacoepidemiol. Drug Saf.* 31, 46–54. doi:10.1002/pds.5323

Bakoyannis, G., and Yiannoutsos, C. T. (2015). Impact of and correction for outcome misclassification in cumulative incidence estimation. *PLOS ONE* 10, e0137454. doi:10.1371/journal.pone.0137454

Carrigan, G., Bradbury, B. D., Brookhart, M. A., Capra, W. B., Chia, V., Rothman, K. J., et al. (2022). External comparator groups derived from real-world data used in support of regulatory decision making: use cases and challenges. *Curr. Epidemiol. Rep.* 9, 326–337. doi:10.1007/s40471-022-00305-9

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models.* 2nd Edition. Chapman and Hall/CRC.

Center for Biologics Evaluation and Research, U.S. Food and Drug Administration (2021). *ABECMA Clinical Review Memo.* 113–114. Available at: https://www.fda.gov/media/147740/download?attachment (Accessed: 03-March-2024)

U.S. Department of Health and Human Services, Food and Drug Administration (2023). *Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. Document No. 2023–02094.* Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products (Accessed: 03-March-2024)

Edwards, J. K., Cole, S. R., Shook-Sa, B. E., Zivich, P. N., Zhang, N., and Lesko, C. R. (2023). When does differential outcome misclassification matter for estimating prevalence? *Epidemiology* 34, 192–200. doi:10.1097/EDE.0000000000001572

Edwards, J. K., Cole, S. R., Troester, M. A., and Richardson, D. B. (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am. J. Epidemiol.* 177, 904–912. doi:10.1093/aje/kws340

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., et al. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* 45, 228–247. doi:10.1016/j.ejca.2008.10.026

Facon, T., Kumar, S. K., Plesner, T., Orlowski, R. Z., Moreau, P., Bahlis, N., et al. (2021). Daratumumab, lenalidomide, and dexamethasone versus lenalidomide and dexamethasone alone in newly diagnosed multiple myeloma (MAIA): overall survival results from a randomised, open-label, phase 3 trial. *Lancet Oncol.* 22, 1582–1596. doi:10.1016/S1470-2045(21)00466-6

Foster, R. B., Lipitz, N. G., Torres, A. Z., and Carson, K. R. (2018). The real-world frequency of 24-hour urine protein electrophoresis (UPEP), serum free light chain

(SFLC), and serum protein electrophoresis (SPEP) testing in patients with multiple myeloma (MM). *Blood* 132, 3536. doi:10.1182/blood-2018-99-113508

Foster, R. B., Tromanhauser, M., Gayer, C., Gonzales, P. K., Maignan, K., Opong, A. S., et al. (2019). Approximating international myeloma working group uniform response criteria to derive response for multiple myeloma (MM) patients using data from electronic health records (EHR). *Blood* 134, 4727. doi:10.1182/blood-2019-127181

Hariton, E., and Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research: study design: randomised controlled trials. *BJOG Int. J. Obstet. Gynaecol.* 125, 1716. doi:10.1111/1471-0528.15199

Hong, S., Schmitt, N., Stone, A., and Denne, J. (2012). Attenuation of treatment effect due to measurement variability in assessment of progression-free survival. *Pharm. Stat.* 11, 394–402. doi:10.1002/pst.1524

Huang Bartlett, C., Mardekian, J., Cotter, M. J., Huang, X., Zhang, Z., Parrinello, C. M., et al. (2020). Concordance of real-world versus conventional progression-free survival from a phase 3 trial of endocrine therapy as first-line treatment for metastatic breast cancer. *PLOS ONE* 15 (4), e0227256. doi:10.1371/journal.pone.0227256

Innes, G. K., Bhondoekhan, F., Lau, B., Gross, A. L., Ng, D. K., and Abraham, A. G. (2021). The measurement error elephant in the room: challenges and solutions to measurement error in Epidemiology. *Epidemiol. Rev.* 43, 94–105. doi:10.1093/epirev/mxab011

Kapetanakis, V., Prawitz, T., Schlichting, M., Ishak, K. J., Phatak, H., Kearney, M., et al. (2019). Assessment-schedule matching in unanchored indirect treatment comparisons of progression-free survival in cancer studies. *Pharmacoeconomics* 37, 1537–1551. doi:10.1007/s40273-019-00831-3

Keogh, R. H., Carroll, R. J., Tooze, J. A., Kirkpatrick, S. I., and Freedman, L. S. (2016). Statistical issues related to dietary intake as the response variable in intervention trials. *Stat. Med.* 35, 4493–4508. doi:10.1002/sim.7011

Kumar, S., Paiva, B., Anderson, K. C., Durie, B., Landgren, O., Moreau, P., et al. (2016). International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. *Lancet Oncol.* 17, e328–e346. doi:10.1016/S1470-2045(16)30206-6

LoCasale, R. J., Pashos, C. L., Gutierrez, B., Dreyer, N. A., Collins, T., Calleja, A., et al. (2021). Bridging the gap between RCTs and RWE through endpoint selection. *Ther. Innov. Regul. Sci.* 55, 90–96. doi:10.1007/s43441-020-00193-5

Lin, K. H., Huynh, L., Yang, X., Zanardo, E., Matay, L., Pinaire, M., et al. (2023). Correlation between Real-World Progression-Free Survival (rwPFS) and Overall Survival (OS) in Patients with Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma (CLL/SLL) Who Relapsed or Are Refractory, Intolerant, Resistant to Treatment with a Bruton's Tyrosine Kinase Inhibitor (BTKi). *Blood* 142, 4652. doi:10.1182/blood-2023-180793

Meinert, C. L. (1996). Clinical trials: the gold standard for evaluation of therapy. *Ophthalmology* 103, 869–870. doi:10.1016/S0161-6420(96)30594-0

Meldrum, M. L. (2000). A brief history of the randomized controlled trial: from oranges and lemons to the gold standard. *Hematol. Oncol. Clin. N. Am. Underst. Clin. Trials* 14, 745–760. doi:10.1016/s0889-8588(05)70309-9

Mhatre, S. K., Machado, R. J. M., Ton, T. G. N., Trinh, H., Mazieres, J., Rittmeyer, A., et al. (2023). Real-world progression-free survival as an endpoint in lung cancer: replicating atezolizumab and docetaxel arms of the OAK trial using real-world data. *Clin. Pharmacol. Ther.* 114 (6), 1313–1322. Epub 2023 Sep 28. PMID: 37696652. doi:10.1002/cpt.3045

Oksen, D., Prince, P., Boutmy, E., Garry, E. M., Ellers-Lenz, B., Estrin, A., et al. (2022). Treatment effectiveness in a rare oncology indication: lessons from an external control cohort study. *Clin. Transl. Sci.* 15, 1990–1998. doi:10.1111/cts.13315

Panageas, K. S., Ben-Porat, L., Dickler, M. N., Chapman, P. B., and Schrag, D. (2007). When you look matters: the effect of assessment schedule on progression-free survival. *J. Natl. Cancer Inst.* 99, 428–432. doi:10.1093/jnci/djk091

Roose, J., Lu, Y., Pierre, A. E., Sawas, A., and Calip, G. S. (2022). Impacts of the COVID-19 pandemic on assessment of M-protein and free light chain levels in patients with multiple myeloma. *Blood* 140, 13191–13193. doi:10.1182/blood-2022-159044

Schmidli, H., Häring, D. A., Thomas, M., Cassidy, A., Weber, S., and Bretz, F. (2020). Beyond randomized clinical trials: use of external controls. *Clin. Pharmacol. Ther.* 107, 806–816. doi:10.1002/cpt.1723

Seeger, J. D., Davis, K. J., Iannacone, M. R., Zhou, W., Dreyer, N., Winterstein, A. G., et al. (2020). Methods for external control groups for single arm trials or long-term uncontrolled extensions to randomized clinical trials. *Pharmacoepidemiol Drug Saf.* 29 (11), 1382–1392. Epub 2020 Oct 4. PMID: 32964514; PMCID: PMC7756307. doi:10.1002/pds.5141

Sibbald, B., and Roland, M. (1998). Understanding controlled trials: why are randomised controlled trials important? *BMJ* 316, 201. doi:10.1136/bmj.316.7126.201

Sondhi, A., Weberpals, J., Yerram, P., Jiang, C., Taylor, M., Samant, M., et al. (2023). A systematic approach towards missing lab data in electronic health records: a case study in non-small cell lung cancer and multiple myeloma. *Syst. Pharmacol.* 12, 1201–1212. doi:10.1002/psp4.12998

Ton, T. G., Pal, N., Trinh, H., Mahrus, S., Bretscher, M. T., Machado, R. J., et al. (2022). Replication of overall survival, progression-free survival, and overall response in chemotherapy arms of non–small cell lung cancer trials using real-world data. *Clin. Cancer Res.* 28 (13), 2844–2853. doi:10.1158/1078-0432.CCR-22-0471

U.S. Department of Health and Human Services, Food and Drug Administration (2018). Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. *Guid. Industry*. Document No. E7-9345. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics (Accessed: 03-March-2024)

Vader, D. T., Mamtani, R., Li, Y., Griffith, S. D., Calip, G. S., and Hubbard, R. A. (2023). Inverse probability of treatment weighting and confounder missingness in electronic health record-based analyses: a comparison of approaches using plasmode simulation. *Epidemiology* 34, 520–530. doi:10.1097/EDE.0000000000001618

Webster-Clark, M., Jonsson Funk, M., and Stürmer, T. (2020). Single-arm trials with external comparators and confounder misclassification: how adjustment can fail. *Med. Care* 58, 1116–1121. doi:10.1097/MLR.0000000000001400

Zhu, J., and Tang, R. (2022). A proper statistical inference framework to compare clinical trial and real-world progression-free survival data. *Stat. Med.* 41, 5738–5752. doi:10.1002/sim.9590