UCLA
UCLA Previously Published Works

Title

Enhancing Safety Signaling: Integrating Clinical Trials and Post-Marketing Adverse Event Reports

Permalink

https://escholarship.org/uc/item/7wb0q5xr

Journal

Manufacturing & Service Operations Management

ISSN

1523-4614

Authors

Bravo, Fernanda
Chen, Lawrence Yunliang
Silberholz, John

Publication Date

2024-03-04

DOI

10.1287/msom.2024.0890

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Enhancing Safety Signaling: Integrating Clinical Trials and Post-Marketing Adverse Event Reports

Fernanda Bravo
Anderson School of Management, UCLA, fernanda.bravo@anderson.ucla.edu

Lawrence Yunliang Chen
Industrial Engineering and Operations Research, University of California, Berkeley, yunliang.chen@berkeley.edu

John Silberholz
Robert H. Smith School of Business, University of Maryland josilber@umd.edu

Problem definition: Adverse drug reactions (ADRs) cause numerous emergency room visits and thousands of deaths annually in the U.S. Although regulators like the Food and Drug Administration (FDA) aim to detect safety issues during clinical trials, these trials are often too small, short, or restrictive to capture rare or delayed ADRs. Consequently, regulators rely on spontaneous adverse event reporting systems (*e.g.*, FAERS) to identify potential safety issues—termed safety signals—by detecting side effects that occur disproportionately among patients taking a particular drug. However, these observational data are subject to selection and reporting biases, which can lead to false signals or missed risks. **Methodology/results:** We propose a Bayesian framework to enhance the hypothesis generation step in safety signaling, offering two methodological advances over standard approaches that rely solely on observational data. First, it integrates pre-approval clinical trial and post-approval observational data. Second, it uses information from common ADRs to estimate and correct bias in rare ADRs. We assess the approach using an analytical model and a large-scale numerical study with real clinical trial and FAERS data. Incorporating clinical trial data substantially reduces signaling costs, and using a common ADR for debiasing yields further gains. This refinement improves decisions for about 1% of ADRs, translating to expected cost savings of 1.5–2.3% (approx. $32–159 million), with performance robust to moderate parameter misspecification. **Managerial implications:** By using existing regulatory data, our approach supports a more efficient and timely hypothesis generation process in post-approval surveillance.

Key words: Health Care Management, Public Policy, Dynamic Programming.

1. Introduction

The safety of pharmaceutical drugs is a major concern for health and industry regulators such as the U.S. Food and Drug Administration (FDA). In the U.S. alone, adverse drug reactions (ADRs) cause thousands of deaths each year, contribute to hospitalizations and longer hospital stays, and result in over a hundred billion dollars in healthcare costs (Lazarou et al. 1998, Richardson et al. 2000, Formica et al. 2018). Identifying safety risks is critical for healthcare providers when weighing drug benefits and harms and for patients making informed treatment choices. Reflecting the importance of timely monitoring, the FDA listed drug safety surveillance as a top priority in 2022 and increased its budget to enhance these efforts (Eglovitch 2021).

Pre-marketing clinical trials are designed to detect key safety concerns—this is the primary focus of Phase I studies and a central goal of Phases II and III (FDA 2018). Safety evidence from these trials, and to a lesser extent from pre-clinical animal studies, forms the basis for the warnings and precautions listed on drug labeling at the time of approval. However, pre-marketing trials are often insufficient for uncovering the full spectrum of a drug's safety profile (Yeleswarapu et al. 2014, Schick et al. 2017). First, their relatively small sample sizes limit the ability to detect rare adverse drug reactions (ADRs), such as those occurring in only 1 in 1,000 patients. Second, the short duration of most trials constrains the observation of side effects that emerge only after prolonged use. Third, restrictive eligibility criteria may exclude patients with comorbidities, polypharmacy, or other relevant characteristics, leading to limited diversity in the study population. As a result, trials may miss ADRs that manifest only in specific subgroups. Together, these limitations underscore the need for robust post-marketing surveillance to identify new, rare, or delayed safety risks.

As a running example, consider Bosutinib, a tyrosine kinase inhibitor approved in the U.S. in 2012 for treating chronic myeloid leukemia (CML) in adults. In pre-approval clinical trials, cardiac failure—a condition where the heart cannot pump sufficient blood to meet the body's needs—was rarely observed. Across four trials, 6 of 544 patients in the treatment group experienced cardiac failure, compared to 1 of 279 in the control group, yielding a non-significant association ($p = 0.43$). Accordingly, the original drug label did not list cardiac failure as a potential adverse effect.

Because of such limitations in clinical trials, the FDA conducts post-marketing drug safety surveillance—known as pharmacovigilance—to monitor the safety of approved products. This ongoing process helps detect new or rare adverse reactions, particularly those emerging from long-term or chronic use, drug interactions, off-label usage, or medication errors. When a causal link is established, the FDA may take regulatory actions, including dosage adjustments, label changes (*e.g.*, adding boxed warnings), or, in rare cases, market withdrawal.

Post-marketing drug surveillance in the U.S. often begins with spontaneous adverse event reports submitted to the FDA by healthcare professionals and patients (voluntarily) or by manufacturers (mandated). These reports are compiled in the FDA's Adverse Event Reporting System (FAERS), which receives over two million entries annually (FDA 2023). Similar systems exist globally, including VigiBase (WHO), EudraVigilance (EMA), and Yellow Card (UK). The FDA continuously analyzes these reports to detect *unusual patterns* in ADR frequency. A key tool is disproportionality analysis (DA), which compares ADR rates among users of a drug for a specific indication versus non-users. Widely used in both regulatory and industry settings, DA supports early detection of potential safety issues, known as *safety signals*. This *hypothesis generation* stage is critical, as it informs which events warrant further investigation. Given their volume and timeliness, spontaneous reports play an outsized role in this process; one study found that over 50% of FDA safety label

changes were based on this data source (Lester et al. 2013). For a broader overview of the FDA surveillance pipeline, see FDA, CDER, CBER (2019) and Ahuja et al. (2021).

Returning to Bosutinib, post-marketing surveillance data indicate potential concern around cardiac failure. Between 2004 and 2020, FAERS recorded cardiac failure as the primary ADR in 52 of 3,515 reports involving Bosutinib (1.5%), compared to 0.4% in the rest of the FAERS database—suggesting a significantly elevated reporting rate ($p < 0.001$). In response, the FDA announced a safety signal for Bosutinib and cardiac failure in the fourth quarter of 2018.

While useful, safety signals generated via disproportionality analysis must be interpreted cautiously, as the underlying observational data are prone to both selection and reporting biases. Selection bias arises when patients who take the drug differ systematically from those who do not. Unlike randomized clinical trials—where treatment arms are balanced—post-market settings often exhibit imbalances. For example, physicians may preferentially prescribe new therapies to sicker patients, who are more likely to experience certain ADRs, thereby inflating the reports for those ADRs even when the drug has no causal effect on them (Lublóy 2014). Additional factors, such as age or socioeconomic status, may also influence treatment uptake and bias safety estimates. Notably, as we discuss in more detail in Section 3, selection bias can simultaneously affect multiple ADRs; for instance, diabetes patients may overreport both renal ADRs and cardiovascular ADRs. Reporting bias occurs when patients taking a particular drug are more or less likely to report an ADR than those not taking the drug. This phenomenon is well-documented, as FAERS is known to suffer from widespread under-reporting (Alatawi and Hansen 2017). Even when two drugs have the same true ADR rate, differences in reporting propensity can create the illusion of disproportionality if one group is more likely to report certain events. A variety of factors can influence reporting patterns, including patient demographics (*e.g.*, gender, age), the length of time a drug has been on the market, and the level of public or regulatory attention an ADR receives (*e.g.*, Dhodapkar et al. (2022), Lee et al. (2023). Since selection and reporting biases tend to affect numerous ADRs simultaneously, our method is structured to take advantage of this shared bias structure.

Understanding the magnitude of selection and reporting biases is critical, as they can lead to misguided regulatory actions with real consequences for patients. For example, Chantix, approved in 2006 to aid smoking cessation, received an FDA boxed warning in 2009 for psychiatric risks following an unusual number of suicide reports in FAERS data (Gann 2011). However, after a seven-year legal battle and $300 million in settlements, Pfizer conducted a large clinical trial—including patients with psychiatric histories—which found no increased risk of depression or suicidal ideation. The results showed that the original concerns were driven by bias in the observational data. The FDA removed the boxed warning in 2016 (Grover 2016), but by then, many smokers may have avoided Chantix due to perceived risks, potentially missing the opportunity for effective treatment.

To address the limitations of disproportionality analysis, regulators such as the FDA incorporate additional methods into their post-market surveillance pipeline to validate safety concerns and filter out spurious signals. One such method involves using electronic health records (EHR) from partner institutions for causal inference analyses, controlling key patient covariates to reduce selection and reporting bias. Even with the FDA's Sentinel Initiative, establishing causality can take months, and 10–40% of signals take over three years to resolve (Beninger and Murray 2021, Dhodapkar et al. 2022). Another option is to mandate post-approval clinical studies, but these are even more costly and take longer to complete. Because spurious signals are expensive to investigate and can divert resources from genuine safety concerns, regulators must be selective in deciding which signals warrant further follow-up. Improving the quality of hypothesis generation would enable more efficient allocation of regulatory resources.

In this work, we focus on the hypothesis generation step in post-approval drug surveillance—specifically, identifying ADRs that merit further investigation. A central challenge is the trade-off between data sources: while clinical trial data is generally unbiased, it often lacks the size or duration to detect rare ADRs; observational data is more abundant but susceptible to selection and reporting biases. To bridge this gap, we propose using information from a common ADR—one well-characterized in both clinical trials and observational data—to debias observational estimates for a rare ADR. The key advantage is that the drug's effect on the common ADR can be estimated with relatively high precision from trials.

To illustrate, consider Bosutinib and the rare ADR of cardiac failure, using gastrointestinal (GI) disorders as the common ADR.[1] Clinical trials were inconclusive regarding cardiac failure, but FAERS data indicated a potential association—albeit with possible bias. In contrast, both data sources yield precise and consistent estimates for GI side effects: a 3.97-fold odds increase in trials and 3.06-fold in FAERS (both $p < 0.001$). This consistency may lend confidence to the overall reliability of the observational data, supporting a signal for cardiac failure. Our method formalizes this logic, using cross-ADR information to debias effect size estimates. When signals are identified, they trigger follow-up investigations and potential regulatory action—as was the case when the FDA added cardiac failure to Bosutinib's label in October 2019.

Framework and contributions: We take an *analytical modeling approach* to studying post-approval safety surveillance, modeling the regulator's labeling and signaling decision for a specific drug-ADR pair within a two-period Bayesian Decision Analysis framework. The regulator decides whether to label the ADR at approval time or to monitor it in post-approval surveillance. The labeling at approval serves as a filtering mechanism, defining the population of ADRs for which post-approval signal generation is relevant. Raising a post-approval safety signal triggers follow-up investigations and potential regulatory actions such as labeling changes. We assign costs to signaling

errors that can be based on error magnitude—Type I error costs capture wasted resources and possible labeling mistakes from spurious safety signals, while Type II error costs capture patient harm from delayed or missed signals. Within this framework, we make the following contributions:

- We propose a "cross-ADR debiasing approach" that combines pre-approval clinical trial data with post-marketing surveillance data across multiple ADRs. Assuming selection and reporting biases affect both common and rare ADRs, we model their relationship via a bivariate normal distribution estimated from historical data within a therapeutic class. Using Bayesian methods, we combine these data sources to produce a debiased estimate of the drug's effect on the rare ADR, and we optimize the safety signaling decision across a general class of loss functions. To our knowledge, this is the first modeling of the signaling decision within a Bayesian Decision Analysis framework.
- While access to cross-ADR information is always statistically beneficial in our setting, implementation of the cross-ADR debiasing approach is not costless (see details in Section 4.2). We analytically identify the conditions under which cross-ADR debiasing generates the greatest value, assuming that model parameters—including the statistical relationship between bias in common and rare ADRs—are correctly estimated from historical data. Our analysis quantifies the value of that information and identifies the factors that amplify or diminish its benefits. These results provide guidance on when leveraging information from related ADRs is most valuable and, consequently, where investments in implementing more sophisticated signal-detection and debiasing methods are likely to yield the greatest returns.
- We validate our method using clinical trial data from clinicaltrials.gov (40,176 trials, 1990–2019) and FAERS data (11M reports, 2004–2020). Across 28,000+ drug-ADR pairs for which clinical trial generalization bias is of limited concern and the drug's effect on the common ADR is well estimated, our approach improves signaling decisions over benchmarks that either ignore bias or fail to leverage common ADRs. Notably, by using a common ADR to debias, our approach can further improve signaling decisions for about 1% of ADRs, reducing expected costs by 1.5–2.3% ($32–159 million) over combining clinical trial and observational data to debias. Performance remains robust under moderate parameter misspecification.

Our proposed approach presents a promising opportunity to improve safety signaling for regulators, as it relies on their existing and readily available data sources related to clinical trial outcomes and post-marketing safety reporting.

2. Related Literature

Our work relates to multiple streams of literature, both in pharmacovigilance and beyond.

A large pharmacovigilance literature focuses on detecting potential safety signals from spontaneous reporting data collected post-approval (*e.g.*, the FAERS database). As noted in Section 1,

these data are valuable for their scale but limited by sparse patient covariates and selection and reporting biases. The most common detection methods use disproportionality analysis (DA) to identify ADRs reported at disproportionately high rates among patients taking the drug of interest.

These methods have been broadly used by regulators worldwide, including by the Medicines and Healthcare Products Regulatory Agency in the United Kingdom, the Netherlands Pharmacovigilance Centre, the World Health Organization, and the FDA (Zannas et al. 2014). As reviewed in Dias et al. (2015), many DA methods have been proposed, including simpler approaches such as Proportional Reporting Ratios (PRR) and Reporting Odds Ratios (ROR) introduced in Section 3 as well as more sophisticated Bayesian approaches such as the Empirical Bayes Geometrical Mean (EBGM), the Multi-item Gamma-Poisson Shrinker (MGPS), the Information Component (IC), and Bayesian Confidence Propagation Neural Networks (BCPNN). Signals are typically raised based on the measure of occurrence (*e.g.*, PRR, IC), its statistical significance (*e.g.*, chi-squared test), and the absolute number of reports. For instance, van Puijenbroek et al. (2002) reports a positive association whenever $PRR - 1.96SE > 1$ and $IC - 2SD > 0$. Signaling thresholds in the literature are typically ad hoc. To our knowledge, we are the first paper to apply a Bayesian decision analysis approach to safety signaling. This introduces a more structured and economically grounded decision-making framework and provides theoretical justification for certain signaling approaches in the literature (for instance, we show that signaling based on a z-score threshold is equivalent to minimizing expected cost within a specific family of loss functions).

Supporting these data mining procedures is extensive work addressing common challenges in spontaneous reporting data, such as duplicate reports (Lin and Lo 2013) and misspelled and nonstandard drug names (Wong et al. 2015). Solutions include pipelines for duplicate detection and vocabulary standardization (Banda et al. 2016) as well as platforms providing post-processed drug side effects data for easier automation (*e.g.*, Tatonetti et al. (2012)). Despite the extensive development and use of DA methods with spontaneous reporting data, we are not aware of any works that debias the DA estimate of a drug's effect on a rare ADR using the DA estimate for a more common ADR, as we propose. Our strategy could enhance any existing DA methods in practice.

There is substantial literature on using longitudinal data sources for postmarketing surveillance, such as electronic health record (EHR) data and medical claims data. These data are richer than spontaneous reporting databases, containing detailed patient covariates (EHR data) and longitudinal records of drug exposures and adverse events. Leveraging this richness, numerous methods incorporating temporal trends and time-varying covariates have been developed to investigate safety signals (Arnaud et al. 2017), with evidence from the U.S. (Ryan et al. 2013) and Europe (Schuemie et al. 2013) showing strong performance. In the operations literature, Ahuja et al. (2021) use logistic regression and matching designs to assess drug–ADR associations, demonstrating that

the inaccurate black-box warning on rosiglitazone could have been avoided. Since such methods rely on careful cohort construction, covariate selection, and matching, they are mainly used to confirm rather than automatically generate new signals; for instance, the FDA has used EHR data accessed via its Sentinel system this way. Our work complements this stream by focusing on generating higher-quality signals for later evaluation with longitudinal databases.

Though much of the literature for longitudinal databases of health records relates to confirming safety signals, there is also significant literature on generating new safety signals (our task of "hypothesis generation") from a variety of observational data sources, including EHR data. Sequential statistical testing (SST) can be used to monitor ADRs using longitudinal health record data, with classical approaches such as the sequential probability ratio test (SPRT) (Wald 1945) and maximized SPRT (Brown et al. 2007). An exemplary example of this type of approach within the operations literature can be found in Goh et al. (2015), where the authors propose an algorithm that can be used for continuous monitoring of multiple potentially correlated ADRs; their approach provides Type I error control, is robust to distributional assumptions, and takes into account the temporal sequence of using the drug and experiencing the ADR. Data sources beyond spontaneous reporting databases or longitudinal health records have also been considered to generate new safety signals—these include molecular fingerprints (Vilar et al. 2011), medical literature findings (Xu and Wang 2014, Shetty and Dalal 2011), and social media data (Liu and Chen 2015, Yeleswarapu et al. 2014, Li et al. 2020). In this literature, authors often propose to obtain signals from multiple data sources and combine them to obtain more confidence before raising a final safety signal (Li et al. 2020, Harpaz et al. 2013). Given its scale, spontaneous reporting data is likely to remain an important data source for safety signaling, either as a standalone source or in combination with others, motivating our focus on improving signaling quality by removing bias.

Methodologically, our work draws on causal inference research combining observational and experimental data, including reweighting RCT data (*e.g.*, Hartman et al. (2015)), extrapolating RCT estimates (*e.g.*, Kallus et al. (2018)), and using Bayesian hierarchical models (*e.g.*, Jackson et al. (2006)). Closest to our approach is Athey et al. (2020), which combines secondary-outcome information from RCTs with primary- and secondary-outcome data from observational studies; we build on these statistical methods, adopting a decision-making perspective to offer managerial insights on their benefits. To our knowledge, we are the first to apply these methods in pharmacovigilance to debias spontaneous reports, demonstrating their value in our case study.

Our approach connects to classical Bayesian decision analysis, specifically the one-sided discrimination problem studied by Marshall and Olkin (1968), where a binary classification decision is made under an unknown continuous parameter. We establish structural results showing the optimal policy has a threshold (z-score) form for power loss functions, extend the analysis to a multi-period

setting, and provide new managerial insights by comparing learning approaches and identifying when our method is especially valuable.

3. Model

Consider two ADRs $i \in 1,2$ for a given drug, where ADR 1 is common and ADR 2 is rare. The decision maker aims to assess whether the drug increases the rate of ADR 2 across the full patient population treated by the drug, potentially leveraging information about ADR 1 to improve inference. Let μ_i denote the effect size of the drug on ADR $i \in 1,2$, where $\mu_i > 0$ ($\mu_i < 0$) indicates that the drug increases (decreases) the rate of ADR i. The regulator's objective is to determine whether $\mu_2 > 0$. Define p_i^j as the probability that a randomly selected patient with the relevant indication experiences ADR i under treatment ($j = 1$) or without treatment ($j = 0$). The commonly used log odds ratio effect size is then given by:

$$\mu_i = ln\Big(\frac{p_i^1/(1-p_i^1)}{p_i^0/(1-p_i^0)}\Big). \tag{1}$$

Other effect sizes for binary outcomes, such as the log risk ratio ($\mu_i = ln(p_i^1/p_i^0)$), can also be used. Note that the odds ratio and risk ratio are comparable for rare events.

Although regulators perform surveillance continuously through time, in this work we consider a stylized 2-period decision framework. The first period corresponds to the drug approval time (A) and the second period corresponds to a post-approval or signaling time (S), which can be years after the original approval date. At approval time, the regulator must decide whether to *label* ADR 2—include it in the drug label—based on clinical trial evidence alone. If the regulator chooses not to label the ADR based on clinical trial evidence, then they monitor the drug and collect observational data on the occurrence of adverse drug reactions in the general population. At signaling time, the regulator analyzes the observational data and decides whether to *signal* ADR 2 based on updated evidence. Labeling or raising a safety signal for an ADR that is not caused by the drug is costly (*e.g.*, conducting further investigations, preventing patients from using the drug, etc.) while missing (or delaying the identification of) safety concerns can harm the population.

The central quantity for signaling decisions is the rare ADR effect size μ_2, estimated from both clinical trial and observational data. In this section and in Section 4, we assume complete knowledge of all other model *parameters* and *distributions*—specifically, that the regulator has accurate estimates of key quantities such as the means and variances of bias terms, obtained from historical data on other drugs in the same therapeutic class. Under this setting, the true bias realizations remain unobserved but are treated as random variables with known distributional parameters. While in reality model parameters are subject to uncertainty, we take the common approach of first studying our strategy under a perfect-information setting to obtain an upper bound on its

practical benefits and to derive insights into where it may offer the greatest benefit. We then try to obtain a realistic view of the possible benefits in the presence of estimation errors via our parameter estimation process in Section 5.2 and our parameter misspecification analysis in Section 5.5.

In this full-knowledge setting, we present a statistical model (Section 3.1) and decision framework (Section 3.2) for safety signaling using our proposed "cross-ADR" debiasing approach. All model notation is summarized in Table A1.

3.1. Statistical model

At approval time, the regulator makes labeling decisions based on the clinical trial outcomes. Clinical trials estimate the new drug's effect size from the observed rate of the ADR in the treatment and control arms, for instance by computing the effect size in eq. (1) with p_i^j replaced by $\hat{p}_i^j$, the observed proportion of patients in the clinical trial taking ($j=1$) or not taking ($j=0$) the drug of interest and experiencing ADR i. We assume clinical trial effect size estimates are given by:

$$\begin{aligned} \hat{\mu}_1^C &= \mu_1 + b_1^C + \epsilon_1^C \\ \hat{\mu}_2^C &= \mu_2 + b_2^C + \epsilon_2^C, \end{aligned} \tag{2}$$

where b_i^C are generalization biases due to external validity concerns such as over-restrictive inclusion/exclusion criteria in the clinical trial (a canonical example would be trials excluding patients taking some drug that turns out to have a drug-drug interaction with our drug of interest), or the presence of heterogeneous drug effects (see e-companion for a discussion of how HTEs are linked to clinical trial generalization bias). We assume the b_i^C values are bivariate normally distributed with known means μ_i^{bC} and covariance matrix $\mathbf{\Sigma}^{bC}$. Clinical trial sampling errors ϵ_i^C are also assumed to be bivariate normally distributed with mean zero and known covariance matrix $\mathbf{\Sigma}^{sC}$. The distributional assumptions for sampling error follow from the central limit theorem, while the distributional assumptions for generalization bias are made for analytical convenience. Specifically,

$$\mathbf{\Sigma}^{bC} = \begin{bmatrix} (\sigma_1^{bC})^2 & \rho^{bC}\sigma_1^{bC}\sigma_2^{bC} \\ \rho^{bC}\sigma_1^{bC}\sigma_2^{bC} & (\sigma_2^{bC})^2 \end{bmatrix} \text{ and } \mathbf{\Sigma}^{sC} = \frac{1}{n^C}\begin{bmatrix} (\sigma_1)^2 & \rho^{sC}\sigma_1\sigma_2 \\ \rho^{sC}\sigma_1\sigma_2 & (\sigma_2)^2 \end{bmatrix}.$$

The correlations ρ^{bC} and ρ^{sC} capture the generalization bias and sample correlation link between the common and rare ADRs, respectively, while σ_i^{bC} captures the variability of generalization bias from one drug to another, σ_i captures the difficulty of measuring the effect size for ADR i, and n^C is the total enrollment in the clinical trial. For simplicity in modeling, we assume that the variance of the effect size for each ADR, σ_i^2, is a fixed constant independent of the effect size μ_i. This is a common assumption when modeling effect sizes for binary outcomes, and is appropriate for ADRs with a known baseline rate in the population.[2] Because of the generally small sample sizes n^C and possible external validity concerns of the clinical trials, they provide a noisy estimate of the true effect size of interest μ_2.

Once the drug is on the market, the regulator monitors its safety in the general population by collecting observational data on adverse reactions through spontaneous reporting systems such as FAERS. We model this by assuming that the regulator has $n^O > n^C$ adverse event reports for patients taking the drug of interest. We assume that the regulator performs disproportionality analysis using the (log) *reporting odds ratio* (ROR), a common measure in the pharmacovigilance literature (as detailed in Section 2, numerous other measures have been proposed). The ROR is estimated by computing the effect size in eq. (1) with p_i^j replaced by $\hat{p}_i^j$, the proportion of safety reports from patients with the indication(s) of interest that mention ADR i, broken down by patients taking the drug of interest ($j = 1$) or not ($j = 0$).

We now arrive at the key mechanism of observational bias considered in this work. Consider two patient populations that have different patterns of reporting ADRs. For instance, diabetic patients experience elevated rates of renal and cardiovascular ADRs. As a result, these ADRs may represent a higher proportion of their FAERS reports than they would for non-diabetics. Meanwhile, ADRs not as impacted by diabetes, such as musculoskeletal and lymphatic ADRs, will consist of a smaller proportion of all reports for diabetics than for non-diabetics (since the proportions must sum to 100%). If diabetics are overrepresented in observational data, we may see a positive ROR for side effects like renal and cardiovascular ADRs and a negative ROR for musculoskeletal and lymphatic ADRs, even if the drug has no effect; the opposite is true if diabetics are underrepresented. These are selection biases in the observational data. Across a set of drugs pulling from the diabetic vs. non-diabetic population at different rates, we expect some biases to be positively correlated (*e.g.*, renal with cardiovascular or musculoskeletal with lymphatic) and others to be negatively correlated (*e.g.*, renal with musculoskeletal).

We assume that the reporting odds ratios take the following form:

$$\begin{aligned} \hat{\mu}_1^O &= \mu_1 + b_1^O + \epsilon_1^O \\ \hat{\mu}_2^O &= \mu_2 + b_2^O + \epsilon_2^O, \end{aligned} \tag{3}$$

where b_i^O captures reporting and selection biases; we assume they are bivariate normally distributed with known mean $\boldsymbol{\mu}^{bO}$ and covariance matrix $\boldsymbol{\Sigma}^{bO}$.[3] By the Central Limit Theorem, the sampling errors ϵ_i^O are assumed to be bivariate normally distributed with zero mean and known covariance matrix $\boldsymbol{\Sigma}^{sO}$. Specifically,

$$\boldsymbol{\Sigma}^{bO} = \begin{bmatrix} (\sigma_1^{bO})^2 & \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} \\ \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} & (\sigma_2^{bO})^2 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}^{sO} = \frac{1}{n^O}\begin{bmatrix} (\sigma_1)^2 & \rho^{sO}\sigma_1\sigma_2 \\ \rho^{sO}\sigma_1\sigma_2 & (\sigma_2)^2 \end{bmatrix},$$

where n^O is the observational data size (number of safety reports in FAERS for the focal drug). Note that collecting a large volume of observational data shrinks sampling errors but has no effect on the bias distribution. The observational data are bivariate normally distributed, and correlations

ρ^{bO} and ρ^{sO} capture the reporting and selection bias and sampling noise link between the common and rare ADRs in the observational data. Recalling that the core idea of our proposed approach is to learn about the bias of the rare ADR (b_2^O) based on an accurate estimate of the bias of the common ADR (b_1^O), it is clear that ρ^{bO}, which controls the statistical relationship between the two biases, is a key parameter of our model. We empirically observe, for several drug classes, that the bias correlation is approximately bivariate normally distributed with a range of non-zero correlation values (see Appendix B.1), supporting our debiasing approach.

Starting from a vague prior $\boldsymbol{\mu}$, Bayesian inference at approval yields the clinical trial outcome as the effect size estimate, potentially subject to generalization bias:

$$\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}^C \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}^C - \boldsymbol{\mu}^{bC}, \boldsymbol{\Sigma}^{bC} + \boldsymbol{\Sigma}^{sC}\right). \tag{4}$$

For ease of notation, we denote the mean effect size estimated at approval time as $\hat{\boldsymbol{\mu}}^A := \hat{\boldsymbol{\mu}}^C - \boldsymbol{\mu}^{bC}$. The effect size estimate at signaling time S (after observing clinical trial results and collecting observational data) can be written as

$$\boldsymbol{\mu}|\hat{\boldsymbol{\mu}}^O, \hat{\boldsymbol{\mu}}^C \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}^S, \boldsymbol{\Sigma}^S\right), \tag{5}$$

where $\hat{\boldsymbol{\mu}}^S$ is the posterior mean effect size and $\boldsymbol{\Sigma}^S$ is the posterior covariance matrix after incorporating the observational data. Note that $\hat{\boldsymbol{\mu}}^S$ and $\boldsymbol{\Sigma}^S$ both depend on $\boldsymbol{\Sigma}^{sC}$, $\boldsymbol{\Sigma}^{bC}$, $\boldsymbol{\Sigma}^{bO}$, and $\boldsymbol{\Sigma}^{sO}$, and that $\hat{\boldsymbol{\mu}}^S$ additionally depends on $\hat{\boldsymbol{\mu}}^C$, $\hat{\boldsymbol{\mu}}^O$, $\boldsymbol{\mu}^{bC}$, and $\boldsymbol{\mu}^{bO}$; the precise calculations to determine the quantities follow standard Bayesian updating methodology (details provided in e-companion).

3.1.1. Structural assumptions. Our analysis focuses on the case where ADR 1 is common (say, occurring at a 10% rate) and ADR 2 is rare (say, occurring at a 0.1% rate in the population). For the broadly used log odds ratio and log risk ratio effect sizes, this implies that σ_1^2 is much smaller than σ_2^2—it is much easier to estimate the effect size for the common ADR than for the rare one. We capture this formally via Assumption 1.

Assumption 1. *The common ADR effect size can be precisely estimated in clinical trials; we analyze our base model in the limit as* $\frac{\sigma_1^2}{n^C} \to 0$ *and* $\frac{\sigma_1^2}{n^O} \to 0$.

We relax this assumption in the numerical study (Section 5). For our base model, we also assume that external validity minimally affects the common ADR, so trial findings for it can be reliably generalized to the broader population.

Assumption 2. *Generalization bias for the common ADR is small; that is,* $(\sigma_1^{bC})^2 \to 0$

This assumption is valid for drugs that do not exhibit heterogeneous effects across subpopulations. We make Assumption 2 in the main model both to focus on a more parsimonious model and to better match our setup in the numerical analysis (see Section 5). That said, we show numerically in Section 5 that clinical trial data frequently support this assumption (*i.e.*, we find many drugs for which gastrointestinal disorders, our choice of common ADR, exhibit low generalization bias). We relax Assumption 2 and show that most of the analytical results are preserved under a more general data model (see Section D in e-companion).

Under Assumptions 1 and 2, our data model for the common ADR considerably simplifies from the general models in (2) and (3):

$$\hat{\mu}_1^C = \mu_1, \text{ and } \hat{\mu}_1^O = \mu_1 + b_1^O,$$

meaning $\hat{\mu}_1^O - \hat{\mu}_1^C$ provides an accurate estimate for the common ADR's observational data bias b_1^O. This, in turn, provides an improved bias estimate for the rare ADR:

$$b_2^O | \{b_1^O = x\} \sim \mathcal{N}(\mu_2^{bO} + \frac{\sigma_2^{bO}\rho^{bO}}{\sigma_1^{bO}} \cdot (x - \mu_1^{bO}), (1 - (\rho^{bO})^2) \cdot (\sigma_2^{bO})^2),$$

which can be used to improve the estimate of the rare ADR effect size μ_2 even in the extreme case where the rare ADR is not observed in the clinical trial (a case we will explore numerically in Section 5). In this case, our cross-ADR debiasing approach adjusts the estimate $\hat{\mu}_2^O$ down by an additional $\frac{\sigma_2^{bO}\rho^{bO}}{\sigma_1^{bO}} \cdot (x - \mu_1^{bO})$.

3.2. Decision framework

We develop a stylized two-period safety signaling model to identify safety concerns about a rare ADR. At approval time (denoted "A" in our model notation), the regulator can choose to *label-immediately* and include a warning for the rare ADR in the drug label based on the clinical trial evidence, or to *wait-and-see* and not label at approval time. If the regulator chooses to wait-and-see, they collect (biased) observational data and make a post-approval signaling decision based on the most updated evidence (*i.e.*, combining clinical trial and observational data) at the signaling time (denoted "S" in our model notation). In practice, the regulator decides to raise a safety signal whenever there is sufficient evidence. When a safety signal is raised, it triggers a series of time-consuming and costly investigations to confirm the signal and subsequently establish the causal relationship between the drug and the adverse drug reaction.[4] If there is insufficient evidence to raise a safety signal, the regulator will continue monitoring and may signal at a later time. However, in our two-period model, we assume that signaling occurs only once. Therefore, if a signal is not raised in the current period, it will not be raised in the future. Figure A1 shows the timeline, the data sources, and the regulator's actions in each period.

A key cost in labeling and signaling decisions is the cost of mistakes. We assume the regulator wishes to label (or signal) the rare ADR if its effect size exceeds θ_b, wishes not to label (or signal) if it is below $\theta_a \leq \theta_b$, and is indifferent between θ_a and θ_b. For notational simplicity, we assume $\theta_a \leq 0 \leq \theta_b$, though results trivially extend to cases where $0 < \theta_a \leq \theta_b$ or $\theta_a \leq \theta_b < 0$. The two error types and their costs are interpreted as follows:

A **false negative (Type II error)** occurs when we do not label (or signal) the rare ADR despite $\mu_2 > \theta_b$. We denote the type II error cost associated with failing to raise a safety concern at the approval and signaling time by $T_{\text{II}}^A(\mu_2)$ and $T_{\text{II}}^S(\mu_2)$, respectively. These costs capture any potential health losses (patients' health impact due to doctors prescribing without all the drug safety information) associated with delaying (or even missing) attributing the ADR to the drug.

A **false positive (Type I error)** occurs when we label (or signal) the rare ADR for the drug even though $\mu_2 < \theta_a$. As above, we incur a penalty $T_{\text{I}}^A(\mu_2)$ when we choose to label in the first period and a penalty $T_{\text{I}}^S(\mu_2)$ if we signal in the second period. The cost $T_{\text{I}}^A(\mu_2)$ captures the impact of labeling the ADR at approval time on doctors' and patients' decisions to use the drug (*e.g.*, patients opting out). The cost $T_{\text{I}}^S(\mu_2)$, on the other hand, captures the costs of resources associated with follow-up investigations triggered by signaling the ADR and potential downstream errors (*e.g.*, erroneously concluding that the drug is toxic after performing further investigations).

To capture these costs, we consider a general family of monotone loss functions with finite expected losses whenever μ_2 is normally distributed:

Definition 1 (Monotone loss). Type I error costs $T_{\text{I}}^A(\mu_2)$ and $T_{\text{I}}^S(\mu_2)$ are (weakly) monotone decreasing, are $o(e^{\epsilon\mu_2^2})$ for any $\epsilon > 0$ as $\mu_2 \to -\infty$, take value 0 for $\mu_2 \geq \theta_a$, and assign some cost to errors otherwise ($T_{\text{I}}^A(\mu_2),\ T_{\text{I}}^S(\mu_2) > 0$ for $\mu_2 < \theta_a$). Type II error costs $T_{\text{II}}^A(\mu_2)$ and $T_{\text{II}}^S(\mu_2)$ are (weakly) monotone increasing, are $o(e^{\epsilon\mu_2^2})$ for any $\epsilon > 0$ as $\mu_2 \to \infty$, take value 0 for $\mu_2 \leq \theta_b$, and assign some cost to errors otherwise ($T_{\text{II}}^A(\mu_2),\ T_{\text{II}}^S(\mu_2) > 0$ for $\mu_2 > \theta_b$).

Note that the above definition is general and can capture settings where the decision-maker has an indifference band around zero, *i.e.*, no cost of error is incurred for $\mu_2 \in [\theta_a, \theta_b]$ and $\theta_a < 0 < \theta_b$. For ease of exposition in the remainder of the work, we consider the case $\theta_a = \theta_b = 0$; however, all our results remain valid in the general case stated in Definition 1.

Monotone loss functions satisfy that larger errors lead to larger costs, yet they may have discontinuities. An important special case of our monotone loss function is the subfamily of power loss functions first introduced by Britney and Winkler (1974); specifically,

Definition 2 (μ^k-loss function). A μ^k–loss function is a power loss function of the form

$$T_{\text{I}}(\mu_2) = \begin{cases} 0 & \text{if } \mu_2 \geq 0, \\ M_{\text{I}}|\mu_2|^k & \text{if } \mu_2 < 0, \end{cases} \quad \text{and} \quad T_{\text{II}}(\mu_2) = \begin{cases} 0 & \text{if } \mu_2 \leq 0, \\ M_{\text{II}}|\mu_2|^k & \text{if } \mu_2 > 0, \end{cases}$$

for constants $M_{\text{I}}, M_{\text{II}} > 0$ and $k \geq 0$. We add the superscript "A" or "S" to M_{I} and M_{II} to differentiate the costs at approval or signaling time, respectively.

Note that μ^k-loss functions generalize the concept of a lump sum loss function—the $k = 0$ case assigns a fixed penalty for incorrect labeling or signaling decisions. They also generalize proportional loss functions—the $k = 1$ case assigns a cost that is proportional to the magnitude of the error.

Additionally, we assume that mistakenly labeling an ADR is more expensive than mistakenly signaling the ADR. This is rooted in the observation that erroneously labeling an ADR can have negative consequences on patient decision-making, since doctors rely on the label when making prescribing decisions. In contrast, errors in signaling potential ADRs may result in wasting analyst time and lead to potential downstream errors, but do not directly impact patient decisions. This is because signaling does not induce an immediate change in the drug label.

ASSUMPTION 3. *$T_{\text{I}}^A(\mu_2) \geq T_{\text{I}}^S(\mu_2)$ for all μ_2.*

3.3. Proposed safety signaling approach

To determine the regulator's optimal safety strategy, we compute their expected costs in each period. At approval time, the regulator makes a labeling decision based on the clinical trial data estimate, *i.e.*, the prior effect size estimate $\mu_2|\hat{\boldsymbol{\mu}}^C \sim \mathcal{N}(\hat{\mu}_2^A, (\sigma_2^A)^2)$, where $(\sigma_2^A)^2 := (\sigma_2^{bC})^2 + \frac{1}{n^C}\sigma_2^2$. If the regulator chooses to label the rare ADR at approval time, a type I error cost $T_{\text{I}}^A(\mu_2)$ is incurred, and no further decision needs to be made in the second period. Otherwise, a type II error cost $T_{\text{II}}^A(\mu_2)$ is incurred in the first period, and observational data is gathered until the second decision period. We define the expected loss at approval time as

$$\mathcal{L}_A(\hat{\mu}_2^A, \sigma_2^A, x) := \mathbb{E}_{\mu_2|\hat{\boldsymbol{\mu}}^C \sim \mathcal{N}(\hat{\mu}_2^A, (\sigma_2^A)^2)} \left[T_{\text{I}}^A(\mu_2) \cdot x + T_{\text{II}}^A(\mu_2) \cdot (1 - x) \right], \tag{6}$$

where x takes the value 1 if the regulator chooses to label at approval time, and 0 otherwise.

If the regulator chooses to wait-and-see in the first period, then they make the optimal signaling decision in the second period based on the updated estimate $\mu_2|\hat{\boldsymbol{\mu}}^O, \hat{\boldsymbol{\mu}}^C \sim \mathcal{N}(\hat{\mu}_2^S, (\sigma_2^S)^2)$ (eq. (5)). The optimal expected loss in the second period is

$$\mathcal{L}_S^*(\hat{\mu}_2^S, \sigma_2^S) := \min_{x \in \{0,1\}} \mathbb{E}_{\mu_2|\hat{\boldsymbol{\mu}}^O, \hat{\boldsymbol{\mu}}^C \sim \mathcal{N}(\hat{\mu}_2^S, (\sigma_2^S)^2)} \left[T_{\text{I}}^S(\mu_2) \cdot x + T_{\text{II}}^S(\mu_2) \cdot (1 - x) \right], \tag{7}$$

where x equals 1 when the regulator chooses to signal in the second period and 0 otherwise.

The regulator's optimal labeling decision at approval time is then the one that minimizes the total expected cost in the first and second periods, namely

$$V_A^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \boldsymbol{\Sigma}^{bC}, \boldsymbol{\Sigma}^{sC}, \boldsymbol{\Sigma}^{bO}, \boldsymbol{\Sigma}^{sO}) := \tag{8}$$
$$\min \begin{cases} \mathcal{L}_A(\hat{\mu}_2^A, \sigma_2^A, 1) & \text{Label-immediately} \\ \mathcal{L}_A(\hat{\mu}_2^A, \sigma_2^A, 0) + V_S^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \boldsymbol{\Sigma}^{bC}, \boldsymbol{\Sigma}^{sC}, \boldsymbol{\Sigma}^{bO}, \boldsymbol{\Sigma}^{sO}) & \text{Wait-and-see.} \end{cases}$$

The term $V_S^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \boldsymbol{\Sigma}^{bC}, \boldsymbol{\Sigma}^{sC}, \boldsymbol{\Sigma}^{bO}, \boldsymbol{\Sigma}^{sO})$ corresponds to the second period expected optimal cost associated with the wait-and-see option. Because of the random nature of the observational data, the posterior mean of the effect size after collecting observational data ($\hat{\mu}_2^S$) is an uncertain quantity before collecting that data. Specifically, it follows the so-called *pre-posterior* distribution $\hat{\mu}_2^S|\hat{\mu}_2^C \sim \mathcal{N}(\hat{\mu}_2^A, \nu^2)$, where the variance is given by $\nu^2 := (\sigma_2^A)^2 - (\sigma_2^S)^2$. Hence, the second-period expected optimal cost corresponds to

$$V_S^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \boldsymbol{\Sigma}^{bC}, \boldsymbol{\Sigma}^{sC}, \boldsymbol{\Sigma}^{bO}, \boldsymbol{\Sigma}^{sO}) := \mathbb{E}_{\hat{\mu}_2^S \sim \mathcal{N}(\hat{\mu}_2^A, \nu^2)} \left[\mathcal{L}_S^*(\hat{\mu}_2^S, \sigma_2^S) \right]. \tag{9}$$

Note that the dependency on $\boldsymbol{\Sigma}^{bC}$, $\boldsymbol{\Sigma}^{sC}$, $\boldsymbol{\Sigma}^{bO}$, and $\boldsymbol{\Sigma}^{sO}$ only comes through the variances ν^2 and $(\sigma_2^S)^2$ (see eqs. EC.14 and EC.15 in e-companion) in the optimal expected cost-to-go V_S^*.[5]

3.4. Single-ADR safety signaling approach

For comparison purposes, we define a single-ADR data model in which the regulator collects observational data but does not use the common ADR to debias the rare ADR effect size. Within our framework under Assumption 1, treating the rare ADR in isolation is captured by performing Bayesian updates with modified covariance matrices $\tilde{\boldsymbol{\Sigma}}^{bC} = \begin{bmatrix} (\sigma_1^{bC})^2 & 0 \\ 0 & (\sigma_2^{bC})^2 \end{bmatrix}$ and $\tilde{\boldsymbol{\Sigma}}^{bO} = \begin{bmatrix} (\sigma_1^{bO})^2 & 0 \\ 0 & (\sigma_2^{bO})^2 \end{bmatrix}$. This corresponds to a misspecified data model in which correlations between ADRs are assumed to be zero, though they may not be in reality. As in our proposed approach, all other parameters (*e.g.*, μ_2^{bC}, σ_2^{bC}, μ_2^{bO}, σ_2^{bO}) are assumed known and correctly estimated, so the sole difference from our approach lies in the treatment of ρ^{bO}.

The removal of the off-diagonal terms in the covariance matrices indicates that information about the common ADR cannot be used to learn about the rare ADR.[6] Thus, combining observational and clinical trial data, the Bayesian update for the single-ADR approach yields a posterior effect size estimate with variance $(\tilde{\sigma}_2^S)^2$ (see eqs. EC.14 and EC.16 in e-companion). Of note, the single-ADR approach posterior estimates are less precise than the estimates from the cross-ADR approach, *i.e.*, $(\tilde{\sigma}_2^S)^2 \geq (\sigma_2^S)^2$. Under Assumptions 1–2, this inequality is strict whenever the true $\rho^{bO} \neq 0$.

We assume the single-ADR regulator makes labeling and signaling decisions using the decision framework described in Section 3.2 but using the single-ADR data model above. Intuitively, using this misspecified model—*i.e.*, not using the common ADR to debias the rare ADR effect size—should result in higher costs. We formally show this in the next section and employ the single-ADR model as a benchmark to determine when our cross-ADR debiasing approach is most valuable.

4. Analytical Results

We start by characterizing the structure of the optimal labeling and signaling decisions and show they are of a threshold form in the clinical trial effect size estimate and posterior mean effect

size, $\hat{\mu}_2^A$ and $\hat{\mu}_2^S$, respectively. We then perform comparative statics to identify when our debiasing approach is most beneficial, using the single-ADR approach as a benchmark. For readability we only state the dependency of V_A^* and V_S^* on $\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}$, $\boldsymbol{\Sigma}^{bC}$, $\boldsymbol{\Sigma}^{sC}$, $\boldsymbol{\Sigma}^{bO}$, $\boldsymbol{\Sigma}^{sO}$ when strictly needed. All proofs are included in e-companion.

4.1. Optimal policy

We first characterize the second-period signaling policy structure.

LEMMA 1. *Under monotone loss functions, the second-period optimal signaling policy has a threshold structure on the posterior mean, $\hat{\mu}_2^S$. That is, there exists some unique threshold t_S^*, such that it is optimal to signal whenever $\hat{\mu}_2^S \geq t_S^*$ and not to signal otherwise. Furthermore, t_S^* only depends on the posterior standard deviation σ_2^S and loss function $T_{\mathrm{I}}^S(\mu_2)$ and $T_{\mathrm{II}}^S(\mu_2)$.*

For the special case of μ^k–loss functions, the optimal signaling policy corresponds to a z-score on the posterior distribution (a property that does not hold for general monotone loss functions).

LEMMA 2. *Under μ^k-loss functions $T_{\mathrm{I}}^S(\mu_2)$ and $T_{\mathrm{II}}^S(\mu_2)$ with the same power k, the optimal signaling threshold in the second period corresponds to a z-score. That is, there exists z^* such that the signaling threshold takes the form $t_S^* = z^*\sigma_2^S$, where z^* depends only on loss function parameters M_{I}^S, M_{II}^S, and k.*

In the case of $k=0$, the trade-off captured by the lump-sum loss function resembles the classic newsvendor problem with the overage cost equal to M_{I}^S and underage cost equal to M_{II}^S. The optimal threshold is $z^* = \Phi^{-1}(M_{\mathrm{I}}^S/(M_{\mathrm{I}}^S + M_{\mathrm{II}}^S))$ and, as expected, we signal more aggressively as the cost ratio $M_{\mathrm{II}}^S/M_{\mathrm{I}}^S$ increases. As established in the proof of Lemma 2, when $k>0$, the optimal threshold z^* can be obtained via 1-dimensional root-finding of a strictly monotone decreasing function $f(z) := M_{\mathrm{I}}^S \int_{-\infty}^{-z} \phi(v)(-(v+z))^k dv - M_{\mathrm{II}}^S \int_{-z}^{\infty} \phi(v)(v+z)^k dv$.

We next characterize the first-period labeling policy structure. Proposition 1 shows that the optimal labeling decision in the first period also follows a threshold policy, but on the approval time effect size estimate $\hat{\mu}_2^A$.

PROPOSITION 1. *The optimal labeling policy at approval time is a threshold policy in $\hat{\mu}_2^A$. That is, there exists a unique threshold t_A^* such that it is optimal to wait and see for $\hat{\mu}_2^A \leq t_A^*$ and to label immediately otherwise. Furthermore, with fixed approval-time information $\hat{\mu}_2^A$ and σ_2^A, t_A^* is strictly decreasing in the posterior standard deviation σ_2^S.*

The result above shows that sophisticated learning schemes that reduce posterior variance tend to delay the labeling decision and choose to wait-and-see more than less sophisticated approaches. Hence, the optimal thresholds for the single-ADR approach (denoted as $\tilde{t}_A^*$) and our more sophisticated debiasing approach satisfy $\tilde{t}_A^* \leq t_A^*$; the inequality is strict whenever $\rho^{bO} > 0$.

In general, a single-ADR and a sophisticated learning approach may differ in their labeling decision for intermediate values of $\hat{\mu}_2^A$, but make the same labeling decisions otherwise. That is, when there is little evidence to support the occurrence of the rare ADR in the clinical trial (small $\hat{\mu}_2^A$), both approaches choose to collect additional information. For intermediate $\hat{\mu}_2^A$, the single-ADR approach chooses to label, while the sophisticated approach prefers to wait-and-see. When there is significant evidence of the ADR occurrence from clinical trials (large $\hat{\mu}_2^A$), both approaches choose to label immediately at approval time.

4.2. Benefit of cross-ADR debiasing approach

While access to cross-ADR information is statistically beneficial in our setting, implementing the cross-ADR approach is not cost-free in practice. First, it requires the time and energy of the analyst to make several modeling judgments, such as the selection of common ADRs and which historical data is considered similar enough to be used to estimate various model parameters. Next, analyst effort is often required to confidently link toxicities between clinical trial data and FAERS data (Painter et al. 2026). This effort must be repeated under the cross-ADR model, and may be even more involved due to the fact that the common ADR will often be selected to include a range of different related ADRs. Thirdly, estimating the cross-ADR model may require integrating entirely new data sources (*e.g.*, individual-level patient data from clinical trials), depending on the assumptions of the model (see Section 6 for details), potentially introducing significant data acquisition and cleaning costs for the analyst. Consequently, adoption depends not only on whether the cross-ADR approach is superior, but also on whether the resulting performance gains justify the associated implementation and operating costs. We therefore next characterize the benefits of debiasing using a common ADR and conduct comparative statics to identify the settings in which the cross-ADR approach provides the greatest advantage over the single-ADR model.

We define the benefit as the reduction in optimal expected cost relative to the (lower information) single-ADR approach:

$$B := V_A^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \tilde{\boldsymbol{\Sigma}}^{bC}, \boldsymbol{\Sigma}^{sC}, \tilde{\boldsymbol{\Sigma}}^{bO}, \boldsymbol{\Sigma}^{sO}) - V_A^*(\hat{\boldsymbol{\mu}}^A, \boldsymbol{\mu}^{bO}, \boldsymbol{\Sigma}^{bC}, \boldsymbol{\Sigma}^{sC}, \boldsymbol{\Sigma}^{bO}, \boldsymbol{\Sigma}^{sO}), \tag{10}$$

where $\tilde{\boldsymbol{\Sigma}}^{bC}$ and $\tilde{\boldsymbol{\Sigma}}^{bO}$ differ in the off-diagonal terms relative to $\boldsymbol{\Sigma}^{bC}$ and $\boldsymbol{\Sigma}^{bO}$, linking the rare and common ADRs with observational data biases (see Section 3.4). Recall that while the single-ADR approach may also incorporate observational data to update the effect size estimate, it does not leverage information from the common ADR. Hence, the benefit captures the relative value of learning from a common ADR in addition to combining clinical trial and observational data. Notably, when both approaches decide to wait-and-see at approval time, the benefit coincides with the *expected value of sample information* (EVSI) in Bayesian decision analysis (Raiffa and Schlaifer 1961), which simply measures the expected value of making decisions with a richer information set. The following theorem identifies the conditions under which our approach is most beneficial.

THEOREM 1. *Under Assumptions 1 and 2, the benefit B is non-negative and satisfies the comparative statics in Table 1.*

Table 1 Comparative statics of the benefit

	Parameter	Monotone Loss	μ^k–Loss (Power Loss)
Clinical trial	$\hat{\mu}_1^A$	constant	constant
	$\hat{\mu}_2^A$	unimodal	unimodal
	σ_2	indeterminate	indeterminate
	n^C	indeterminate	indeterminate
	σ_2^{bC}	indeterminate	indeterminate
Observational	μ_i^{bO}	constant	constant
	$\lvert\rho^{bO}\rvert$	increasing	increasing
	σ_1^{bO}	constant	constant
	σ_2^{bO}	indeterminate	unimodal
	n^O	indeterminate	unimodal for $k>1$; increasing for $k\leq 1$ and $\hat{\mu}_2^C z^* \leq 0$

Note: Benefit is, under Assumptions 1 and 2, independent of parameters ρ^{sC}, ρ^{sO}, and ρ^{bC}. The term "indeterminate" means that the benefit does not have a unique behavior with respect to the focal parameter (*i.e.*, it can be increasing, unimodular, bimodal, etc.) depending on the setting.

To understand the comparative statics, we first summarize intuitive behaviors before exploring unexpected findings.

Intuitive behaviors: We first discuss parameters ρ^{bO}, σ_1^{bO}, μ_1^{bO}, and $\hat{\mu}_1^A$, which all play a role in how our proposed approach learns from the common ADR. Intuitively, we learn the direction and magnitude of the bias in the common ADR and then use that information to de-bias the observational data for the rare ADR. Under Assumptions 1 and 2, the clinical trial perfectly measures the effect of the drug on the common ADR (*i.e.*, effect size estimate $\hat{\mu}_1^A$ equals the true effect size μ_1), so the z-score of the selection/reporting bias for the common ADR can be perfectly measured via the normalized difference between the observational and clinical trial effect size estimates, $(\hat{\mu}_1^O - \mu_1^{bO} - \hat{\mu}_1^A)/\sigma_1^{bO}$. Since we precisely learn the direction and magnitude of the bias for the common ADR regardless of the values of $\hat{\mu}_1^A$, μ_1^{bO}, and σ_1^{bO}, they play no role in the comparative statics of our proposed approach (they play a more nuanced role once either Assumption 1 or 2 is relaxed; see Theorem EC.4 in e-companion). A stronger correlation between the common and rare ADR's biases means that our approach can learn more from the common ADR, so an increasing benefit in $|\rho^{bO}|$ is an intuitive result, as well.

Next, we explore the approval time effect size estimate for the rare ADR, $\hat{\mu}_2^A := \hat{\mu}_2^C - \mu_2^{bC}$. If $\hat{\mu}_2^A$ is very large, then we have already seen that both our approach and the single-ADR approach label immediately, resulting in no benefit of our approach. If, on the other hand, $\hat{\mu}_2^A$ is very small (*e.g.*, satisfying $\hat{\mu}_2^A/\sigma_2^A \ll 0$, showing a high likelihood of a protective effect of the drug on the rare ADR, even when taking into account the possibility of generalization bias in the trial), then both

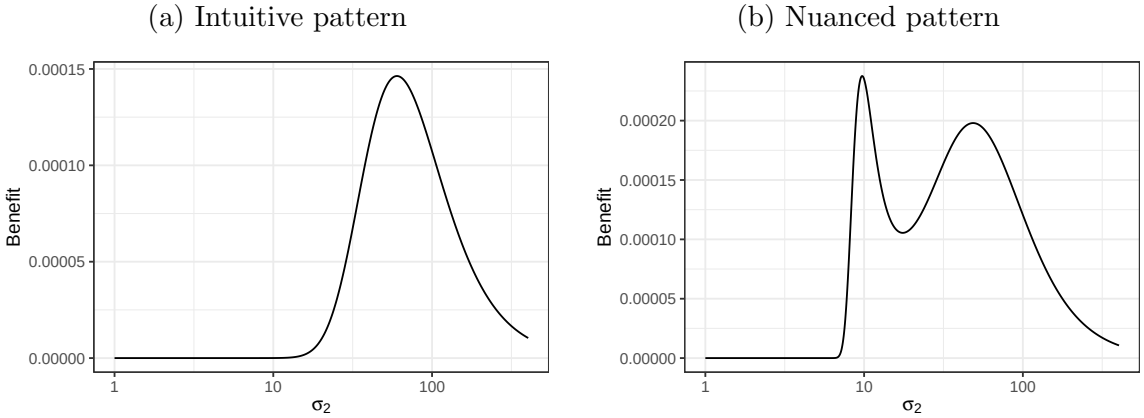
approaches will wait-and-see. The observational data is unlikely to cause us to signal under either approach, given how strongly the clinical trial results indicated a protective effect, meaning the benefit of the proposed approach will be quite small. We conclude that the unimodal (up-down) benefit in $\hat{\mu}_2^A$ reported in Theorem 1 is quite natural—intermediate clinical trial effect size estimates provide the most opportunity for the proposed approach to improve the signaling decision. Note that up-down behavior in $\hat{\mu}_2^A$ translates into an up-down pattern in $\hat{\mu}_2^C$ and μ_2^{bC}.

Lastly, we discuss $\underline{\mu_2^{bO} \text{ and } \sigma_2^{bO}}$, the mean and uncertainty in the amount of selection/reporting bias present in the rare ADR's observational data, under μ^k-loss functions. We subtract out the bias mean before performing debiasing, so μ_2^{bO} has no effect on expected loss when all parameters are known. Our proposed approach provides benefit over the single-ADR approach by de-biasing the observational data for the rare ADR, so of course it provides limited value when σ_2^{bO} is small (there is little-to-no bias to remove). In contrast, when σ_2^{bO} is very large, then the observational data will be of limited value regardless of whether our strategy is used or not—there is simply too much bias to effectively learn from the observational data. In this scenario, the Bayesian update in eq. (5) will put very little weight on the observational data compared to the clinical trial data, meaning the single-ADR approach and our approach will make near-identical decisions. In light of this, the unimodal (up-down) benefit in σ_2^{bO} is again quite natural—intermediate amounts of bias in the observational data for the rare ADR provide the most opportunity for the proposed approach to improve the signaling decision. We note that under more general monotone loss functions, however, the unimodal behavior no longer must hold, as we illustrate in Figure EC.2 in e-companion.

Surprising or nuanced insights: The comparative statics with respect to the rare ADR's rareness $\underline{\sigma_2}$ are of great interest—how rare should an ADR be to maximally benefit from our proposed approach? Intuitively, we may again expect that the rare ADR should be rare, but not too rare, to maximally benefit. When the rare ADR is exceptionally rare (σ_2 is very large), then the clinical trial and observational data sampling errors are large compared to the magnitude of the bias, which means de-biasing provides limited benefit.[7] On the other hand, when the "rare" ADR is actually quite common (σ_2 is small), we can accurately estimate the drug's impact on the "rare" ADR in clinical trials (so long as the generalization error σ_2^{bC} is not too large), resulting in both approaches yielding similar benefit. We therefore might intuitively expect intermediate values of σ_2 to yield the most benefit of our approach, since our ADR is rare enough that the clinical trial does not tell the whole story, but common enough that sampling errors are not too large.

While this first-order intuition of a unimodal (up-down) relationship between ADR rareness σ_2 and the benefit of our proposed approach often holds (see, *e.g.*, Figure 1, left pane), it is not always the case. As we can see in the right pane of Figure 1, there can instead be a bimodal relationship between σ_2 and the benefit of the proposed approach, and the largest benefit can accrue for a

Figure 1 **Examples of indeterminate benefit patterns for μ^k-loss function with $k = 0$.**



Note: The plots show the second-period cost difference $\tilde{V}_S^* - V_S^*$ (this is the relevant portion of the benefit whenever $M_{\text{II}}^A \approx 0$, meaning that it is generally optimal to wait-and-see in the first period). Here, $M_{\text{I}}^S = 1$, $M_{\text{II}}^S = 0.07$, $k = 0$, $n^C = 200$, $n^O = 600$, $\sigma_2^{bO} = 3$, $\rho^{bO} = 0.2$, $\sigma_2^{bC} = 0.3$ (left) and 0.1 (right), and $\hat{\mu}_2^A = 0.3$ (left) and 1 (right). The remaining parameters do not affect the benefit due to Theorem 1 ($\hat{\mu}_1^C$, μ_1^{bC}, μ_i^{bO} and σ_1^{bO}) or Assumptions 1–2 (ρ^{sC}, ρ^{sO}, and ρ^{bC}).

rather common ADR (in the figure, the benefit spike at $\sigma_2 \approx 9.5$ corresponds to an ADR that occurs at a rate of roughly 4.6%). This seemingly counterintuitive finding, a sizable benefit for the proposed approach for a very common "rare" ADR, can be explained by the choice of loss function. Note in the Figure 1 example that our cost function has unbalanced signaling costs $M_{\text{I}}^S \gg M_{\text{II}}^S$, corresponding to a critical z-score $z^* = 1.51$. Trials with $\sigma_2 \approx 9.5$ correspond to situations where the clinical trial z-score is close to the signaling decision boundary—our clinical trial with $\hat{\mu}_2^A = 1$ would have z-score 2.72 with $\sigma_2 = 5$, 1.47 with $\sigma_2 = 9.5$, and 0.71 with $\sigma_2 = 20$. Thus, the spike in benefit with $\sigma_2 \approx 9.5$ can be interpreted as capturing clinical trial results that are precisely estimated (due to a small σ_2) yet close to the signaling decision boundary,[8] causing the observational data to act as a "tiebreaker" in determining whether regulators should signal or not. Since our proposed approach does a better job of learning from observational data than the single-ADR approach, it yields a particularly large benefit over the competitor in this regime. From a regulatory standpoint, this unexpected boundary effect suggests that our debiasing approach provides benefits not only for rare ADRs but also for relatively common ADRs that are challenging to assess in clinical trials due to their borderline nature.

A similar story emerges when considering the impacts of generalization bias noise $\underline{\sigma_2^{bC}}$ and clinical trial sample size $\underline{n^C}$ on the benefit of our proposed approach versus the single-ADR approach. Both of these parameters have a monotone relationship with $(\sigma_2^A)^2$, the variance of our estimator at approval time. Thus, while we might intuitively expect a unimodal (up-down) comparative static under μ^k-loss functions, we can also observe bimodal behavior due to this "tiebreaker effect," with an extra spike in benefit corresponding to cases where the selected σ_2^{bC} or n^C value causes

the clinical trial z-score to fall close to the signaling decision boundary. Examples of these more complicated relationships are provided in Figure EC.3 in e-companion and demonstrate that even quite large clinical trials could plausibly benefit from our approach.

Finally, we observe an interesting pattern of the benefit with respect to the sample size of observational data $\underline{n^O}$. Intuitively, we may expect that more observational data is always beneficial—a small n^O will cause so much sampling error in the observational data that the data will receive limited weight in the Bayesian update and therefore our debiasing strategy will have limited effect, while a large n^O will improve the signal-to-noise ratio of observational data, allowing our strategy to shine. While we see the expected monotone relationship for μ^k-loss functions for some parameter regimes ($k \leq 1,\ \hat{\mu}_2^A z^* \leq 0$),[9] we instead see a unimodal (up-down) comparative static with diminishing benefit for large n^O when using μ^k-loss functions that heavily penalize large mistakes ($k > 1$). The reason for the diminishing benefits when $k > 1$ is that there is more benefit to improving estimation precision when σ_2^S is large (n^O is small), since a large σ_2^S indicates bigger errors and a loss function with a large k disproportionately penalizes such errors. To see this, consider the expected signaling loss under a μ^k-loss function with $k = 2$, $M_{\text{I}}^S = M_{\text{II}}^S = 1$ (critical z-score of 0 for signaling), and a posterior effect size of $\hat{\mu}_2^S = 2$. The expected loss of the optimal action (signaling) improves by 0.42 (from 35.94 to 35.52) when improving the posterior variance $(\sigma_2^S)^2$ from 100 to 99. Meanwhile, it has more modest improvements of 0.26 (from 1.62 to 1.37) when improving the variance from 10 to 9 and of 0.006 (from 0.006 to 0) when improving the variance from 1 to 0. Viewed through this lens, an up-down comparative static for n^O when $k > 1$ makes sense—with intermediate n^O, we collect enough observational data for our proposed approach to improve the quality of our statistical estimator compared to the single-ADR approach, and we avoid collecting so much observational data that both approaches have precise estimates and therefore result in small gaps in the expected loss. From a policymaker's standpoint, this underscores the importance of carefully selecting the loss function, taking into account the specific disease and potential adverse reactions. Notably, with loss functions that disproportionately penalize large errors ($k > 1$), the cross-ADR debiasing approach can return the largest benefits with a moderate amount of observational data, making it especially appealing for less frequently reported ADRs.

5. Numerical Study of Proposed Approach

We now evaluate the practical benefits of our approach using large-scale real-world data from FAERS and clinical trials. We compare our method to benchmark approaches and assess its robustness to parameter misspecification.

5.1. Data

To estimate the model parameters, we consider the following two datasets:

Spontaneous reporting data. Our spontaneous reporting dataset consists of all FAERS reports from 2004 through May 2020 (approximately 26 million reports). We adopt the data cleaning and deduplication approach detailed in Banda et al. (2016), using both Natural Language Processing (NLP) tools as well as hand labeling to standardize drug and indication names.

For each FAERS report, we extract the primary drug ID (identified by RxNorm codes (Nelson et al. 2011)), the primary indication for which the patient was taking this drug (identified by Medical Dictionary for Regulatory Activities (MedDRA v22.1) Preferred Term (PT)), and the ADR(s) experienced by the patient (identified by MedDRA PT terms). Note that the MedDRA dictionary used for indications and ADRs is a hierarchy; PT terms are quite specific (there are approximately 80,000 PT terms), but each PT is nested within a primary System Organ Class (SOC) which is broader (there are 27 SOCs). The resulting FAERS data consists of approximately 11 million FAERS reports, including 2,754 unique primary drugs and 18,324 unique ADRs.

Clinical trials data. The clinical trial data, extracted from ClinicalTrials.gov, covers 40,176 studies from 1990–2019. For each study, we extract the condition (*i.e.*, therapeutic indication), intervention drug(s) in each arm (control and treatment), patient counts per arm, and ADR occurrences. We use NLP tools to map drug names to RxNorm and ADR names to MedDRA PT terms to align with FAERS coding. Ultimately, the clinical trial data includes 1,615 unique drugs and 5,262 unique ADRs, for a total of 403,843 unique drug-ADR pairs.

5.2. Subset identification and model estimation

We use gastrointestinal toxicities—common events occurring in approximately 20% of trial patients—as the common ADR, given their suitability for precise estimation in both clinical trials and spontaneous reports. We next outline how we select relevant drug-ADR pairs and estimate the model parameters from Section 3.

5.2.1. Estimating drug effects. For a given drug, we may have a series of RCTs testing the drug against some control and seek to combine the RCT data to obtain a more precise estimate of the effect of the drug on some ADR of interest; we do this to obtain our clinical trial estimates $\hat{\mu}_i^C$ and σ_i^C in the first period of our framework. Recall the log odds ratio defined in (1), which measures the effect of a drug on the rate of an ADR. In each trial j, we can use the contingency table (number of patients in the treatment and control arm who did and did not experience the ADR) to obtain an estimate $\hat{\mu}^{(j)}$ of the log odds ratio of experiencing our ADR with the selected drug, along with its standard error $\sigma^{(j)}$. To combine these RCT-specific values into a single estimate of the drug's effect on the ADR, we use random-effects meta-analysis (Borenstein et al. 2021). Formally, we assume each trial's treatment effect takes the form

$$\hat{\mu}^{(j)} = \mu + \delta^{(j)} + \epsilon^{(j)}, \tag{11}$$

where μ is the average effect of the drug on the ADR, $\delta^{(j)} \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$ is a random effect capturing differences in the effect from one RCT population to another, and $\epsilon^{(j)}$ are sampling errors. Random-effects meta-analysis delivers a combined estimate $\hat{\mu}^C$ of the drug effect, along with standard error σ^C and estimate T of the heterogeneity parameter τ. Random-effects meta-analysis of clinical trial data can be used to estimate the effects of drugs on common ADRs (*e.g.*, ADRs grouped at the MedDRA SOC level), but will have limited or no information about a rare ADR that was not observed in any clinical trial.

We use FAERS data to obtain our observational data estimates $\hat{\mu}_i^O$ and σ_i^O, where the contingency table of reports with/without the drug and with/without the ADR yields the (log) reporting odds ratio $\hat{\mu}^O$ and its standard error σ^O.

5.2.2. Subset selection. We apply filters to focus on drug-ADR pairs where our proposed approach and estimation procedure are well-motivated. We enforce the following requirements:

1. *Possible safety issue*: We focus on drug-ADRs occurring at least once in the FAERS database.
2. *Limited clinical trial data for rare ADR*: We limit to drug-ADR pairs where the ADR never appeared in any clinical trial arms testing the drug.[10]
3. *Precisely estimated common ADR with limited generalizability concerns*: In the spirit of Assumptions 1–2, we focus on studying drug-ADRs where our common ADR (gastrointestinal toxicity) is precisely estimated with limited clinical trial generalizability concerns. To operationalize this, we define the common ADR to be "well estimated" for a given drug-indication pair if: 1) the sampling variances of the clinical trial and observational data are both small ($(\sigma_1^C)^2, (\sigma_1^O)^2 \leq 0.1$), and 2) there are multiple clinical trials of the drug for that indication, with no signs of effect size heterogeneity across the selected patient populations of the trials (the random-effects meta-analysis of the trial results has heterogeneity parameter estimate $\tau = 0$). We only use data from drug-indications that are well-estimated for the common ADR when estimating the bias distribution in Section 5.2.3, and we only apply our cross-ADR approach for a given drug-rare ADR pair if at least one drug-indication has a "well-estimated" common ADR for that drug and the drug's effect on both the common ADR and the SOC containing the rare ADR is consistent across indications for that drug.[11]
4. *Sufficient historical data*: To ensure sufficient data when estimating the bias distribution (see Section 5.2.3 for details), we group drugs using the Anatomical Therapeutic Chemical (ATC) classification system,[12] limiting to drugs in a drug class with at least 5 drug-indication pairs for which the common ADR is "well estimated."

Filtering drug-ADR pairs in this way requires matching drugs, indications, and ADRs between the FAERS and clinical trial data, motivating our approach to coding these data sources using

the same medical vocabulary. Despite applying multiple filters, we are left with a large number of drug-ADR pairs that we could consider signaling: 28,431 in total, corresponding to 14 different drugs across 4 different drug classes.

5.2.3. Estimating the bias distribution. Recall that we view the key mechanism of observational bias to be the over- or under-representation of a subpopulation among those who elect to take the drug post-approval. Since post-approval usage patterns may vary from one indication to another for the same drug, we collect data for different drug-indication pairs when estimating the patterns of bias.

Consider some focal drug d and a rare ADR. To estimate the bias distribution, we identify a set of "well-estimated" drug-indication pairs $D := \{(d'_1, i_1), (d'_2, i_2), \cdots, (d'_k, i_k)\}$ within the same drug therapeutic class. Thus, Criterion 3 implies that a pair (d', i) satisfies: (i) drug d' is similar to focal drug d (they are in the same drug class); (ii) we have precise estimates of common ADR effect sizes for (d', i), so sampling error correlations are not a major concern; and (iii) we have limited clinical trial generalization concerns for (d', i), as identified by random-effects meta-analysis of multiple clinical trials. Under these stringent requirements, we expect the correlation between $\hat{\mu}_1^O - \hat{\mu}_1^C$ and $\hat{\mu}_2^O - \hat{\mu}_2^C$ to be largely due to observational bias correlation for drug-indication pairs in the set D.

We propose to estimate $\mathbf{b}^O$ by studying the distribution of $\hat{\boldsymbol{\mu}}^O - \hat{\boldsymbol{\mu}}^C$ for "well estimated" drug-indication pairs in the same drug class. Recalling eq. (2) and eq. (3), we can write $\hat{\boldsymbol{\mu}}^O - \hat{\boldsymbol{\mu}}^C = \mathbf{b}^O - \mathbf{b}^C + \boldsymbol{\epsilon}^O - \boldsymbol{\epsilon}^C$. We approximate this distribution as

$$\hat{\boldsymbol{\mu}}^O - \hat{\boldsymbol{\mu}}^C \overset{\text{approx}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1^{bO} \\ \mu_2^{bO} \end{bmatrix}, \begin{bmatrix} (\sigma_1^{bO})^2 + (\sigma_1^C)^2 + (\sigma_1^O)^2 & \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} \\ \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} & (\sigma_2^{bO})^2 + (\sigma_2^C)^2 + (\sigma_2^O)^2 \end{bmatrix} \right). \tag{12}$$

This approximation follows under limited generalization bias, *i.e.*, $\mathbf{b}^C \approx 0$, and under precisely estimated common ADR (sampling error correlations have small magnitude).

Because effect sizes $\hat{\mu}_2^C$ and $\hat{\mu}_2^O$ may not be well estimated in the historical data for a very rare ADR, it may be challenging to confidently estimate bias distribution parameters. To address this, we estimate the bias distribution using not only the rare ADR itself (defined at the MedDRA PT level, *e.g.*, "cardiac arrhythmias") but also data for related ADRs (defined at the MedDRA SOC level, *e.g.*, "cardiac disorders"). As discussed in Section 3.1, it is natural to combine similar toxicities in this way since we expect selection biases to hold across a range of similar toxicities.

To estimate bias distribution parameters μ_1^{bO}, μ_2^{bO}, σ_1^{bO}, σ_2^{bO}, and ρ^{bO} for each drug class and for each SOC that our rare ADR could be a part of, we use maximum likelihood estimation, assigning a regularization penalty to the ρ^{bO} parameter selected via 10-fold cross-validation (see Appendix B.2 for details). This regularization shrinks the estimate of ρ^{bO} towards zero, reducing the likelihood of overestimating the benefit of our approach. We further examine the impact of this shrinkage

through the parameter misspecification analysis (Section E of the e-companion). Ultimately, we estimate bias distribution parameters for five drug classes. We obtain $|\hat{\rho}^{bO}| \geq 0.1$ for 48 out of $26 \cdot 5 = 130$ estimated ρ^{bO} values (see distribution of ρ^{bO} in Figure B3). We also statistically validate the bivariate normality assumption in eq. (12) for $\hat{\boldsymbol{\mu}}^O - \hat{\boldsymbol{\mu}}^C$; see Appendix B.1 for details.

5.3. Benchmarks and performance evaluation

Following Section 5.2, we have fully estimated the parameters of our statistical model from Section 3 and are now prepared to evaluate the performance of our proposed approach versus competitors.

- **Scheme 1: Bias unaware.** In this benchmark approach, safety signaling for a drug-ADR pair is performed entirely based on the reporting odds ratio estimates from the spontaneous reporting data ($\hat{\mu}_2^O$ and $(\sigma_2^O)^2 := (\sigma_2)^2/n^O$), ignoring the existence of bias entirely. This approach is attractive due to its limited data needs (all that is needed is the contingency table from the FAERS database), and it is a common approach in practice (Hoffman et al. 2016).
- **Scheme 2: Single-ADR debiasing.** In this benchmark approach, the regulator is aware of the potential for bias in observational data (they have access to bias mean μ_2^{bO} and variance $(\sigma_2^{bO})^2$), and takes this into account when making signaling decisions—they raise safety signals using the log reporting odds ratio $\hat{\mu}_2^O - \mu_2^{bO}$ and variance $(\sigma_2^O)^2 + (\sigma_2^{bO})^2$. Note that they make no use of the common ADR for debiasing purposes.
- **Scheme 3: Cross-ADR debiasing.** Our proposed approach employs a Bayesian update to obtain the posterior effect size $\hat{\mu}_2^S$ and variance $(\sigma_2^S)^2$, as detailed in Section 3. This extends beyond the single-ADR bias handling benchmark by debiasing using the common ADR.

Ordered from smallest to largest, the variances of the drug-ADR estimates used for safety signaling are 1, 3, and 2. The reason Scheme 1 has the smallest variance is that it completely disregards bias, resulting in overconfidence in its estimates based on observational data.

In practice, regulators often rely on disproportionality methods applied to spontaneous report data (*e.g.*, FAERS). These methods detect signals by comparing observed and expected counts for specific product-event pairs, flagging a signal when the resulting statistic exceeds a predefined threshold. Scheme 1 in our study closely mirrors this real-world practice: it operates solely on spontaneous data and does not adjust for underlying biases. Our proposed approach advances existing methods along two key dimensions: first, by combining clinical trial and observational data, and second, by incorporating information from a common ADR to further enhance inference. Scheme 2 isolates the benefit of the first dimension. It integrates clinical trials with observational data to estimate the effect size of a rare ADR, adjusting for known biases. Scheme 3, our full approach, builds on Scheme 2 by also leveraging information from a common ADR; capturing the second methodological advancement. Comparing Schemes 2 and 3 allows us to quantify the

added value of cross-ADR debiasing in improving post-market surveillance beyond what is achieved through integrating clinical trial data alone.

We compute signaling decisions under each scheme using a monotone loss function defined by cost parameters M_{I} and M_{II} (see eqs. (D.3)–(D.4)). The cost of missing a signal M_{II}, which captures the health effects of excess ADRs caused by the drug, is calibrated at the drug-PT level based on the number of future patients who will use the drug and the severity-adjusted harm of the reaction, with more severe ADRs incurring higher costs. The cost of incorrectly raising a safety signal, M_{I}, is calibrated at the drug level using a target z-score, and we conduct sensitivity analyses with respect to this parameter. Hence, we obtain optimal signaling thresholds (Lemma 1) that vary across PTs for a specific drug—for instance, more severe ADRs are assessed with a more stringent (*i.e.*, lower) threshold. For details on the loss function and cost parameters, see Appendix D. As it is standard in the Bayesian decision analysis literature, we assess the performance loss of schemes 1 and 2 via the *conditional value of sample information* (CVSI) (Raiffa and Schlaifer 1961). Intuitively, CVSI evaluates the quality of a decision by measuring its expected value under the posterior conditioned on the most complete information set—*i.e.*, the posterior that uses common ADR data in our setting—which serves as the benchmark for decision quality. CVSI assumes the most complete posterior is correctly specified; we examine the impact of model misspecification in Section 5.5.

5.4. Results

As expected, for most of the 28,431 drug–ADR pairs, all three schemes yield the same signaling decision. However, for about 7–10% (depending on the target z-score used to calibrate M_{I}), at least one benchmark approach disagrees. Relative to the single-ADR approach, the cross-ADR method changes the decision for roughly 180–250 pairs, and for about 1,800–2,700 pairs compared to the bias-unaware approach, depending on the target z-score.

Table 1 summarizes additional expected costs for each benchmark relative to our cross-ADR debiasing approach for varying z-scores. The bias-unaware (Scheme 1) performs poorly, relying solely on biased observational data and ignoring clinical trial evidence. While the single-ADR debiasing approach significantly improves performance, our approach achieves an additional 1.5–2.3% lower expected cost. Importantly, Scheme 2 is already sophisticated—integrating clinical trial and observational data via Bayesian updating—and the benefit of our approach solely stems from incorporating the common ADR in the debiasing step. In practice, regulators may employ an approach closer to Scheme 1, which relies solely on observational data for safety signals.

Differences in signaling decisions across the approaches are summarized in Table 2. In general, the two benchmark approaches tend to miss more signals than non-signals (see contingency tables for the z-score of 2 in Figure 2b). For illustration, let us consider a z-score of 2. The bias-unaware

Table 1 Additional costs of benchmarks relative to cross-ADR approach

	Δ Expected cost relative to cross-ADR ($million)	
z-score	Scheme 1	Scheme 2
0.5	1,688 (113%)	32 (2.2%)
1	2,246 (90%)	45 (1.8%)
1.5	3,204 (83%)	89 (2.3%)
2	4,859 (86%)	82 (1.5%)
2.5	10,429 (136%)	159 (2.1%)

Note: The z-score parameter captures the balance between type I and type II error costs. Low z-scores imply lower type I error costs, which leads to all approaches signaling more; the opposite pattern occurs for large z-scores.

approach misses 1,496 signals and 707 non-signals, resulting in almost twice as much expected cost as our approach. Furthermore, while the single-ADR approach tends to agree much more with our cross-ADR signaling decisions, it still differs in 137 signals and 63 non-signals, with an estimated expected cost of over $82 million (48% higher expected cost relative to our approach in the 200 drug-PT pairs with different signal recommendations).

Figure 2a shows the distribution of the expected cost ratio for schemes 1 and 2 for the set of drug-ADR pairs with different signaling recommendations (assuming a z-score of 2). The y-axis shows the percentage of drug-ADR pairs with an expected loss ratio of $(1, 1.25]$, $(1.25, 2]$, and > 2. We note that whenever schemes 1 and 3 disagree in their signaling decision, the expected cost ratio tends to be large (greater than 2); this is shown by the size of the darker section of the left-side bar. The superior performance is intuitive, as the bias-unaware scheme places excessive confidence in observational data estimates, which may be heavily affected by reporting and selection biases. Among the drug-ADR pairs where scheme 2 differs from our approach, our approach halves (or better) the expected cost for 25% of those pairs (darker portion of the right-side bar in Figure 2a).

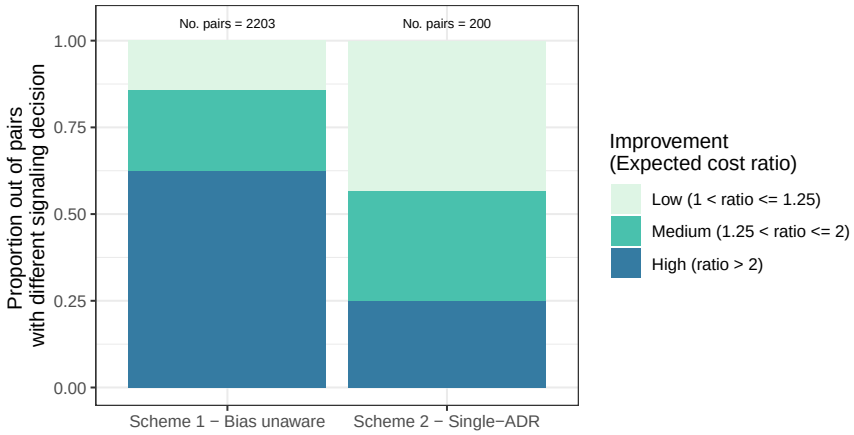
In Appendix B.3, we also present the performance analysis classified by the severity of the ADR (using the severity classification outlined in Lavertu et al. (2021)), the commonness of the ADR (using the reporting frequency of the rare ADR in FAERS), and the drug class. We find that the benefits of the cross-ADR debiasing approach are similar in all severity subgroups. 87% (173 out of 200) of the drug-ADRs where our approach outperforms the single-ADR approach occurred for

Table 2 Signaling decisions relative to cross-ADR approach

		% missed signals			% missed non-signals	
z-score	Cross-ADR signal pairs	Scheme 1	Scheme 2	Cross-ADR non-signal pairs	Scheme 1	Scheme 2
0.5	16,699	12%	0.9%	11,732	6%	0.8%
1	14,221	14%	1.1%	14,210	5%	0.6%
1.5	11,785	15%	1.3%	16,646	4%	0.6%
2	9,565	16%	1.4%	18,866	4%	0.3%
2.5	7,648	16%	1.6%	20,783	3%	0.3%

Figure 2 Comparison of analytical and numerical results.

(a) Expected cost ratio distribution among drug-ADRs with different signaling decisions



(b) Contingency tables

Scheme 1

Scheme 3 \	S	NS
S	28.4%	5.3%
NS	2.5%	63.9%

Scheme 2

Scheme 3 \	S	NS
S	33.2%	0.5%
NS	0.2%	66.1%

Note. We select a z-score of 2 to estimate the drug-specific type I error cost M_{I}. In (a), the percentages are calculated over the entire list of drug-PT pairs (*i.e.*, 28,431).

rare ADRs with fewer than 10 FAERS reports recorded, largely owing to the fact that we focus on ADRs that did not appear in any clinical trial (criterion 2 in Section 5.2.2). The benefit of our proposed approach against the "bias unaware" approach holds broadly across the five drug classes studied, while our performance against the single-ADR debiasing approach is consistent for all but one drug class (nitrogen mustard analogues).

5.5. Sensitivity to parameter mis-specification

We conduct a series of robustness checks to assess the sensitivity of our approach to estimation errors in key model parameters. First, recognizing ρ^{bO} as the key parameter, we analytically compare the performance of a misspecified cross-ADR with that of the single-ADR approach. We show that the cross-ADR is locally superior when the assumed misspecified correlation satisfies either $\underline{\rho}^{bO} \to 0$ or $\underline{\rho}^{bO} \to \rho^{bO}$, for a true $\rho^{bO} > 0$ (the same argument can be made for negative correlation). The latter case corresponds to the assumed parameter values in the analytical modeling. Because strong regularization can drive estimated correlations toward zero, this finding also provides theoretical support for the use of heavily regularized estimates. Details are provided in Section E of the e-companion. Second, we numerically examine the effect of misspecification in all model parameters on performance. We borrow parameter values estimated from the main numerical study for the same set of drug-ADR pairs, moderately perturb them, and compute the resulting expected costs of our approach. For computational tractability, we use a lump sum loss function. Overall, our method consistently outperforms the single-ADR benchmark under moderate misspecification. For details on this analysis and results for all parameters, see Appendix C.

6. Discussion and Conclusions

This work develops a Bayesian framework to improve post-approval drug safety surveillance. Selection and reporting biases make spontaneous reporting system data—the primary source of data in pharmacovigilance—unsuitable for causal inference of the effect of a drug on a rare adverse reaction. Our approach mitigates observational data bias by learning the direction and magnitude of such bias from common ADRs. Our approach's ultimate goal is to enhance the hypothesis generation step in post-approval surveillance, enabling more accurate and expedited signal detection.

We theoretically characterize the conditions under which debiasing using a common ADR offers the greatest advantage for signaling relative to a simpler method that relies solely on clinical trial and observational data for the rare ADR. As expected, the cross-ADR approach yields more conservative labeling decisions at approval, leaving more ADRs under post-approval surveillance. Importantly, we identify when incorporating cross-ADR information provides sufficient benefits to justify the additional costs required for implementation. We complement the analytical results with a numerical comparison against two benchmark methods, demonstrating that our approach can reduce expected signaling costs by an additional 1.5–2.3% beyond what is achievable through debiasing each ADR solely with its clinical trial data.

We believe this work opens promising avenues for future research, most critically around the estimation of model parameters. In Section 5 we considered a simpler setting with a well-estimated common ADR and minimal generalization bias concerns. While this setting simplifies estimation by allowing us to attribute correlations between the common and rare ADR to be observational data bias (see eq. (12) in Section 5.2.3), it does so at the cost of limiting the scenarios where the cross-ADR approach can be applied and also introducing strong assumptions into the estimation procedure (it is difficult to guarantee that generalization bias and sampling errors truly play no role in the correlation). Follow-on work using new data sources to improve estimation would be helpful. First, individual patient data from clinical trials would enable identification of correlated sampling errors ($\rho^{sC} \neq 0$), enabling accurate estimation even when the "common ADR" is somewhat less common. Further, a large-scale observational study operating on covariate-rich electronic health record data could compare unadjusted and covariate-adjusted drug effects to create a library of bias distribution estimates across different ADR pairs for different drug classes and indications; these bias distribution estimates could be used prospectively when applying the cross-ADR approach, and would allow decision makers to separate clinical trial generalization biases from observational data biases. Though this work used stylized modeling assumptions, such as conjugate priors and known parameter values, for analytical tractability, estimation-oriented work could explore a richer set of models that impose fewer structural assumptions and capture parameter uncertainty. Such work could also explore using different subsets of historical data to estimate the model. The current

work could guide all follow-on estimation-oriented studies toward signaling problems where the cross-ADR approach has the most promise.

Beyond parameter estimation, the framework could be extended from the two-period setting to continuous monitoring. Another direction is to optimize the choice of the common ADR to maximize cross-ADR debiasing benefits, or to combine multiple signals from different common ADRs. When several common ADRs are available, rather than relying on any single signal, regulators can view these signals as complementary pieces of evidence. More broadly, integrating this approach with additional data sources, such as post-marketing clinical studies, could be valuable, especially as regulators increasingly prioritize such studies for accelerated approvals (Xu et al. 2021).

Finally, the concept of debiasing observational data with RCT information has applications far beyond safety signaling (for instance, Kallus et al. (2018) apply the idea to education data). Future work could apply the cross-ADR debiasing approach in other domains with multiple outcomes of interest (some easy to measure and some hard to measure).

References

Ahuja V, A Alvarez C, Birge JR, Syverson C (2021) Enhancing regulatory decision making for postmarket drug safety. *Management Science* 67(12):7493–7510.

Alatawi YM, Hansen RA (2017) Empirical estimation of under-reporting in the US Food and Drug Administration adverse event reporting system (FAERS). *Expert opinion on drug safety* 16(7):761–767.

Arnaud M, Bégaud B, Thurin N, Moore N, Pariente A, Salvo F (2017) Methods for safety signal detection in healthcare databases: a literature review. *Expert opinion on drug safety* 16(6):721–732.

Athey S, Chetty R, Imbens G (2020) Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676* .

Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH (2016) A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific data* 3(1):1–11.

Beninger P, Murray M (2021) Review of FDA amendments act section 921 experience in posting data-mining results from the FAERS database. *Clinical Therapeutics* 43(2):380–395.

Borenstein M, Hedges LV, Higgins JP, Rothstein HR (2021) *Introduction to meta-analysis* (John Wiley & Sons).

Britney RR, Winkler RL (1974) Bayesian point estimation and prediction. *Annals of the Institute of Statistical Mathematics* 26(1):15–34.

Brown JS, Kulldorff M, Chan KA, Davis RL, Graham D, Pettus PT, Andrade SE, Raebel MA, Herrinton L, Roblin D, Boudreau D, Smith D, Gurwitz JH, Gunter MJ, Platt R (2007) Early detection of adverse drug events within population-based health networks: application of sequential testing methods. *Pharmacoepidemiology and Drug Safety* 16(12):1275–1284.

ClinicalTrialsgov (2023) Clinical trials. URL `https://clinicaltrials.gov/`.

Dhodapkar MM, Shi X, Ramachandran R, Chen EM, Wallach JD, Ross JS (2022) Characterization and corroboration of safety signals identified from the us food and drug administration adverse event reporting system, 2008-19: cross sectional study. *BMJ* 379.

Dias P, Penedones A, Alves C, F Ribeiro C, B Marques F (2015) The role of disproportionality analysis of pharmacovigilance databases in safety regulatory actions: a systematic review. *Current drug safety* 10(3):234–250.

Eglovitch JS (2021) FDA's FY22 budget priorities: boost inspections, surveillance; avert device shortages. URL https://www.raps.org/resource/fdas-fy22-budget-priorities-boost-inspections-surv.html, Accessed September 22, 2026.

FDA (2018) The Drug Development Process. URL https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process, Accessed September 22, 2026.

FDA (2023) FDA Adverse Events Reporting System public dashboard. URL https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database, Accessed September 22, 2026.

FDA, CDER, CBER (2019) Best Practices in Drug and Biological Product Postmarket Safety Surveillance for FDA Staff - Draft. Technical report, Food and Drug Administration.

Formica D, Sultana J, Cutroneo P, Lucchesi S, Angelica R, Crisafulli S, Ingrasciotta Y, Salvo F, Spina E, Trifirò G (2018) The economic burden of preventable adverse drug reactions: a systematic review of observational studies. *Expert opinion on drug safety* 17(7):681–695.

Gann C (2011) ABC News Medical Unit: FDA should warn smokers of increased danger of chantix. URL https://abcnews.com/Health/chantix-dangers-government-attention-study/story?id=14868835, Accessed September 22, 2026.

Goh J, Bjarnadóttir MV, Bayati M, Zenios SA (2015) Active postmarketing drug surveillance for multiple adverse events. *Operations Research* 63(6):1528–1546.

Grover N (2016) FDA drops black box warning on Pfizer's anti-smoking drug. URL https://www.reuters.com/article/us-pfizer-fda/fda-drops-black-box-warning-on-pfizers-anti-smoking-drug-idUSKBN1452JJ, Accessed September 22, 2026.

Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, Chase HS, Friedman C (2013) Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association* 20(3):413–419.

Hartman E, Grieve R, Ramsahai R, Sekhon JS (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 757–778.

Hoffman KB, Dimbil M, Tatonetti NP, Kyle RF (2016) A pharmacovigilance signaling system based on FDA regulatory action and post-marketing adverse event reports. *Drug safety* 39:561–575.

Jackson C, Best N, Richardson S (2006) Improving ecological inference using individual-level data. *Statistics in medicine* 25(12):2136–2159.

Kallus N, Puli AM, Shalit U (2018) Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 10888–10897.

Lavertu A, Hamamsy T, Altman RB (2021) Quantifying the severity of adverse drug reactions using social media: network analysis. *Journal of Medical Internet Research* 23(10):e27714.

Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA* 279(15):1200–1205.

Lee KM, Rushovich T, Gompers A, Boulicault M, Worthington S, Lockhart JW, Richardson SS (2023) A gender hypothesis of sex disparities in adverse drug events. *Social Science & Medicine* 116385.

Lester J, Neyarapally GA, Lipowski E, Graham CF, Hall M, Pan GD (2013) Evaluation of FDA safety-related drug label changes in 2010. *Pharmacoepidemiology & Drug Safety* 22(3):302–305.

Li Y, Yepes AJ, Xiao C (2020) Combining social media and FDA Adverse Event Reporting System to detect adverse drug reactions. *Drug safety* 43(9):893–903.

Lin WY, Lo CF (2013) Co-training and ensemble based duplicate detection in adverse drug event reporting systems. *2013 IEEE International Conference on Bioinformatics and Biomedicine*, 7–8 (IEEE).

Liu X, Chen H (2015) A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. *Journal of biomedical informatics* 58:268–279.

Lublóy A (2014) Factors affecting the uptake of new medicines: a systematic literature review. *BMC Health Services Research* 14:469.

Marshall AW, Olkin I (1968) A general approach to some screening and classification problems. *Journal of the Royal Statistical Society: Series B (Methodological)* 30(3):407–435.

MedDRA (2023) Medical Dictionary for Regulatory Activities. URL https://www.meddra.org/.

Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R (2011) Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association* 18(4):441–448.

Painter JL, Haguinet F, Bate A (2026) Ctg-db: An ontology-based transformation of clinicaltrials. gov to enable cross-trial drug safety analyses. *arXiv preprint arXiv:2603.15936* .

Raiffa H, Schlaifer R (1961) *Applied statistical decision theory* (Division of Research at the Harvard Business School).

Richardson WC, Berwick DM, Bisgard C, Bristow L, Buck CR (2000) *To err is human: building a safer health system* (National Academy Press).

Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, Reich CG, Schuemie MJ, Madigan D (2013) A comparison of the empirical performance of methods for a risk identification system. *Drug Safety* 36(Suppl 1):S143–S158.

Schick A, Miller KL, Lanthier M, Dal Pan G, Nardinelli C (2017) Evaluation of pre-marketing factors to predict post-marketing boxed warnings and safety withdrawals. *Drug safety* 40(6):497–503.

Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, Innocenti F, Mazzaglia G, Picelli G, van der Lei J, Sturkenboom MCJM (2013) Replication of the OMOP experiment in Europe: Evaluating methods for risk identification in electronic health record databases. *Drug Safety* 36(Suppl 1):S159–S169.

Shetty KD, Dalal SR (2011) Using information mining of the medical literature to improve drug safety. *Journal of the American Medical Informatics Association* 18(5):668–674.

Tatonetti NP, Patrick PY, Daneshjou R, Altman RB (2012) Data-driven prediction of drug effects and interactions. *Science translational medicine* 4(125):125ra31–125ra31.

van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC (2002) A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiology and drug safety* 11(1):3–10.

Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C (2011) Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *Journal of the American Medical Informatics Association* 18(Supplement_1):i73–i80.

Wald A (1945) Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* 16(2):117–186.

Wong CK, Ho SS, Saini B, Hibbs DE, Fois RA (2015) Standardisation of the FAERS database: a systematic approach to manually recoding drug name variants. *Pharmacoepidemiology and drug safety* 24(7):731–737.

Xu L, Zhao H, Petruzzi NC (2021) Inducing compliance with postmarket studies for drugs under FDA's accelerated approval pathway. *Manufacturing & Service Operations Management* 23(1):170–190.

Xu R, Wang Q (2014) Automatic signal extraction, prioritizing and filtering approaches in detecting post-marketing cardiovascular events associated with targeted cancer drugs from the FDA Adverse Event Reporting System (FAERS). *Journal of biomedical informatics* 47:171–177.

Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R (2014) A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC medical informatics and decision making* 14(1):1–16.

Zannas AS, Okuno Y, Doraiswamy PM (2014) Cholinesterase inhibitors and pisa syndrome: a pharmacovigilance study. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 34(3):272–278.

Online Appendix for Enhancing Safety Signaling: Integrating Clinical Trials and Post-Marketing Adverse Event Reports

Fernanda Bravo
Anderson School of Management, UCLA, fernanda.bravo@anderson.ucla.edu

Lawrence Yunliang Chen
Industrial Engineering and Operations Research, University of California, Berkeley, yunliang.chen@berkeley.edu

John Silberholz
Robert H. Smith School of Business, University of Maryland josilber@umd.edu

Endnotes

[1]GI disorders (*e.g.*, nausea, vomiting, and diarrhea) are included in Bosutinib's original label under "Warnings and Precautions" (DailyMed 2023).

[2]For instance, the log odds ratio effect size for ADR i defined in eq. (1) has variance $\frac{\sigma_i^2}{n} = \frac{2}{n}\left(\frac{1}{p_i^0} + \frac{1}{1-p_i^0} + \frac{1}{p_i^1} + \frac{1}{1-p_i^1}\right)$ in a trial with n patients evenly divided between the treatment (index 1) and control (index 0) arms. An ADR with a baseline rate of $p_i^0 = 0.001$ in the control arm will have $\sigma_i = 63.3$ when $\mu_i = 0$ (no effect of the treatment on the ADR rate), will have $\sigma_i = 61.8$ when $\mu_i = 0.1$ (treatment increases the odds of the ADR by roughly 10%), and will have $\sigma_i = 64.9$ when $\mu_i = -0.1$ (treatment reduces the odds of the ADR by roughly 10%).

[3]We assume b_i^O and b_i^C are uncorrelated, as clinical trial bias often results from narrow inclusion criteria, whereas observational bias typically arises from selection or reporting differences. Since these sources are distinct, a significant correlation is unlikely. Our numerical evaluation in Section 5 focuses on the setting where generalization bias is not large, limiting the impact of this modeling choice on the paper's numerical results.

[4] Signal validation is a multi-step process involving detailed case reviews and supporting safety data. The FDA prioritizes signals based on seriousness, novelty, or unexpected frequency, and classifies individual reports as "probable," "possible," or "unlikely." Probable cases trigger a case series review, which considers clinical and demographic details, exposure patterns, and reporting trends. If a signal is confirmed, sponsors are required to conduct follow-up studies (*e.g.*, observational studies, registries), and the FDA evaluates the results to determine causality and guide regulatory actions (Report of CIOMS Working Group VIII 2010).

[5]Despite not being the primary goal of safety surveillance systems, our model could be adapted to guide the removal of ADRs from drug labels. However, Type I and Type II error costs would need to be calibrated accordingly.

[6]Following Assumption 1, the off-diagonal terms of $\mathbf{\Sigma}^{sC}$ and $\mathbf{\Sigma}^{sO}$ are already assumed to be approximately 0.

[7]A bit more formally, it is easy to check that the difference $(\tilde{\sigma}_2^S)^2 - (\sigma_2^S)^2$ between the posterior variances of the single-ADR strategy and our strategy approaches a constant as $\sigma_2 \to \infty$, even though both posterior variances are growing large (scaling with σ_2^2). This implies that $\hat{\mu}_2^S/\sigma_2^S$, the z-score after collecting the observational data, will be

nearly identical for the two strategies when σ_2 is large, meaning that they will almost always make the same signaling decision (see Lemma 2).

[8]More precisely, the peak occurs where $\hat{\mu}_2^A/\sigma_2^S$, the clinical trial effect size normalized by the posterior variance after collecting observational data, is right at the decision boundary. Here we consider a clinical trial that precisely estimates the effect size, so the distinction is minor — with $\sigma_2 = 9.5$, we have a clinical trial z-score $\hat{\mu}_2^A/\sigma_2^A = 1.47$, while we have $\hat{\mu}_2^A/\sigma_2^S = 1.51$.

[9]Notice in the case of $k \leq 1$ that monotone increasing benefit requires $\hat{\mu}_2^A z^* \leq 0$. In fact, the "tiebreaker effect" we saw for σ_2, n^C, and σ_2^{bC} can again play a role when $k \leq 1$ and $\hat{\mu}_2^A z^* > 0$, as can be seen in Figure EC.1 in e-companion. Recall from Section that a spike in benefit can occur when $\hat{\mu}_2^A/\sigma_2^S$ coincides with the critical z-score for the signaling loss function. The increases in n^O correspond to decreases in σ_2^S, and the spike in Figure EC.1 in e-companion at $n^O \approx 10,000$ corresponds to the n^O value at which $\hat{\mu}_2^A/\sigma_2^S$ falls at the decision boundary. Interestingly, the "tiebreaker effect" does not play a role in the comparative statics for n^O when $k > 1$.

[10] We focus on new, extremely rare ADRs for which regulators, lacking trial evidence, initially take a wait-and-see approach and monitor post-market data. While the FDA also monitors ADRs seen in trials (*e.g.*, cardiac failure in bosutinib), we exclude these to avoid assumptions about regulator behavior. Our focal ADRs will not be labeled at approval, so approval-time loss functions T_{I}^A and T_{II}^A need not be estimated. Evaluating debiasing for more common ADRs is left for future work.

[11]Formally: we perform random-effects meta-analysis to combine the $\hat{\mu}_i^C$ values across indications for the drug, and require that this second random-effect meta-analysis also not show signs of heterogeneity (we require heterogeneity parameter estimate $\tau = 0$).

[12]This system defines a hierarchy of drugs; we use level 4 of the hierarchy to obtain chemical subgroups of drugs. We chose to group drugs within a therapeutic class because they are chemically similar, are often studied in similar patient populations (subject to similar inclusion and exclusion criteria in clinical trials), and may exhibit similar patterns of confounding and reporting in observational data.

Appendix A: Notation and supporting material

Table A1 Notation

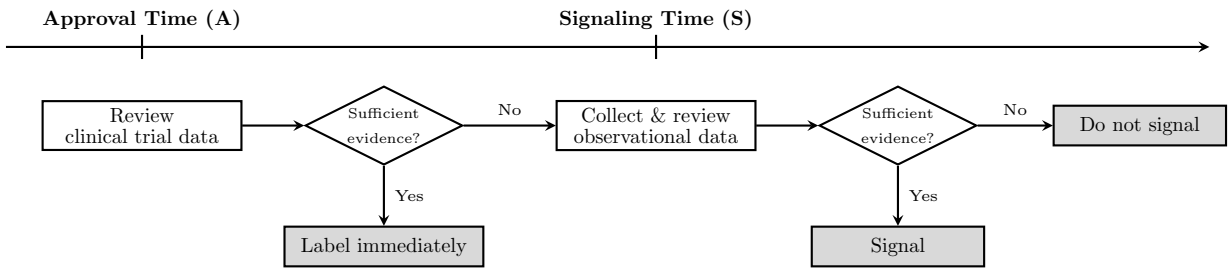
μ_i	true effect size of ADR i
$\hat{\mu}_i^C$ ($\hat{\mu}_i^O$)	effect size estimate in clinical trials (observational data) for ADR i
σ_i	population standard deviation for ADR i
n^C (n^O)	size of the clinical trial (observational data)
ϵ_i^C (ϵ_i^O)	sampling noise of clinical trial (observational data) for ADR i
ρ^{sC} (ρ^{sO})	clinical trial (observational data) sampling correlation between the common and rare ADRs
b_i^C	generalization bias in clinical trials for ADR i
μ_i^{bC}	mean clinical trial generalization bias for ADR i
ρ^{bC}	clinical trial generalization bias correlation between the common and rare ADRs
σ_i^{bC}	standard deviation of clinical trial generalization bias for ADR i
b_i^O	observational data bias for ADR i
μ_i^{bO}	mean observational bias for ADR i
ρ^{bO}	observational data bias correlation between the common and rare ADRs
σ_i^{bO}	standard deviation of observational data bias for ADR i
$\mathbf{\Sigma}^{sC}$ ($\mathbf{\Sigma}^{sO}$)	covariance matrix of clinical trials data (observational data) sampling errors
$\mathbf{\Sigma}^{bC}$ ($\tilde{\mathbf{\Sigma}}^{bC}$)	covariance matrix of generalization bias (single-ADR approach)
$\mathbf{\Sigma}^{bO}$ ($\tilde{\mathbf{\Sigma}}^{bO}$)	covariance matrix of observational data bias (single-ADR approach)
$\hat{\mu}_2^A$	ADR 2 *prior* mean effect size estimate at approval time
σ_2^A	ADR 2 *prior* effect size standard deviation at approval time
$\hat{\mu}_2^S$	ADR 2 *posterior* mean effect size estimate at signaling time
σ_2^S ($\tilde{\sigma}_2^S$)	ADR 2 *posterior* effect size standard deviation at signaling time (single-ADR approach)
$\nu(\tilde{\nu})$	*pre-posterior* standard deviation of $\hat{\mu}_2^S$ (uncertainty at approval time)
$T_{\mathrm{I}}^l(\cdot)$	type I error cost function at time $l \in \{A, S\}$
$T_{\mathrm{II}}^l(\cdot)$	type II error cost function at time $l \in \{A, S\}$
M_{I}^l	type I error cost of μ^k-loss function at time $l \in \{A, S\}$
M_{II}^l	type II error cost of μ^k-loss function at time $l \in \{A, S\}$
k	power of the μ^k-loss function
$\mathcal{L}_l(\cdot)$	expected loss function at time $l \in \{A, S\}$
$\mathcal{L}_S^*(\cdot)$	optimal expected loss function at signaling time
$V_l^*(\cdot)$	optimal expected cost at time $l \in \{A, S\}$
t_l^*	optimal mean effect size threshold for labeling (or signaling) at time $l \in \{A, S\}$
B	benefit or cost difference between the cross-ADR approach and the single-ADR approach

Appendix B: Numerical analysis

B.1. Testing of bias bivariate normality assumption

Recall from eq. (12) that we assume our bias estimates $\Delta\hat{\mathbf{b}} := \hat{\boldsymbol{\mu}}^O - \hat{\boldsymbol{\mu}}^C$ to be approximately bivariate normal under the filters defined in Section 5.2.2. We test this assumption via the Henze-

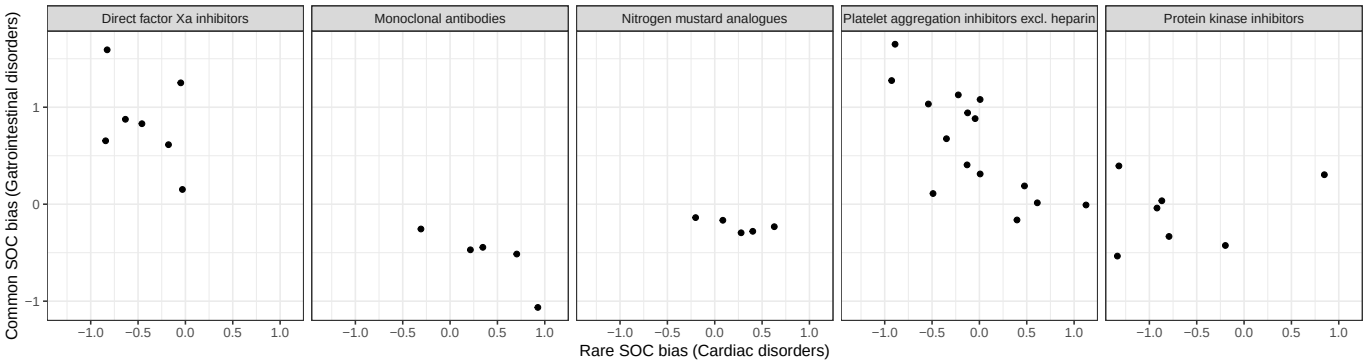
Figure A1 Timeline of data collection and decision framework



Zirkler test applied to bias difference estimates $\Delta\hat{\mathbf{b}}$. The test's p-value can be used to reject the null hypothesis of bivariate normality for the observational bias. Given the small number of drug-indications in each drug class, we can only test bivariate normality for one of our drug classes (Platelet aggregation inhibitors excl. heparin), and we obtain a p-value > 0.05 for all 26 SOCs.

We illustrate the bias data using the cardiac disorders and gastrointestinal SOCs in Figure B2.

Figure B2 Example of SOC-level bias data used to test the bivariate normal assumption



B.2. Maximum-likelihood estimation of bias distribution parameters

As described in Section 5.2.3, the core data used to estimate the bias distribution are clinical trial and observational data for drug-indications of chemically similar drugs (drugs in the same level 4 ATC drug class), filtered according to a set of criteria laid out in Section 5.2.2. For each drug-indication pair (d, i) in our drug class $\mathcal{C}$, we obtain estimates for the SOC of the common ADR (which we will denote with index "1") and of the SOC of the rare ADR (which we will denote with the index "2"). As discussed in Section 5.2.3, the MedDRA SOC term associated with an ADR is a higher-level categorization capturing not only that ADR but also similar ones. For each drug-indication, random-effects meta-analyses of clinical trial results (see Section 5.2.1) yield effect size estimates $\hat{\mu}^C_{(d,i),1}$ and $\hat{\mu}^C_{(d,i),2}$ of the effect of the drug on the incidence of the two SOCs, along with standard errors $\sigma^C_{(d,i),1}$ and $\sigma^C_{(d,i),2}$. Further, the reporting odds ratios from FAERS for each drug-indication pair yields observational data effect size estimates $\hat{\mu}^O_{(d,i),1}$ and $\hat{\mu}^O_{(d,i),2}$ and standard errors $\sigma^O_{(d,i),1}$ and $\sigma^O_{(d,i),2}$.

Following equations (2) and (3), the differences between the observational and clinical trial effect size estimates can be re-expressed in terms of bias and sampling error terms:

$$\hat{\mu}^O_{(d,i),1} - \hat{\mu}^C_{(d,i),1} = b^O_{(d,i),1} - b^C_{(d,i),1} + \epsilon^O_{(d,i),1} - \epsilon^C_{(d,i),1}, \text{ and}$$
$$\hat{\mu}^O_{(d,i),2} - \hat{\mu}^C_{(d,i),2} = b^O_{(d,i),2} - b^C_{(d,i),2} + \epsilon^O_{(d,i),2} - \epsilon^C_{(d,i),2}.$$

Before estimating the parameters of our system, we make two simplifying assumptions motivated by the filters applied to our drug-indication pairs (see Section 5.2.2, criterion 3):

1. Since we parameterize our model using only drug-indication pairs with limited concerns related to the generalizability of clinical trial results, we assume that the biases $b^C_{(d,i),1}, b^C_{(d,i),2} \approx 0$.
2. Since we parameterize our model using only drug-indication pairs where both the clinical trial and observational data effect sizes for the common ADR are well estimated from data (have small standard errors), we assume that the covariance term for our sampling errors are small and that our system can be well approximated by assuming that $\rho^{sO}, \rho^{sC} \approx 0$.[13]

Recall that the bias terms have means μ_1^{bO} and μ_2^{bO}, which we will estimate. Now, we can appeal to the distributions of bias $\mathbf{b^O_{(d,i)}}$ and sampling errors $\boldsymbol{\epsilon}^C_{(d,i)}$ and $\boldsymbol{\epsilon}^O_{(d,i)}$ to obtain the distribution of our effect size differences, which replicates eq. (12):

$$\begin{bmatrix} \hat{\mu}^O_{(d,i),1} - \hat{\mu}^C_{(d,i),1} \\ \hat{\mu}^O_{(d,i),2} - \hat{\mu}^C_{(d,i),2} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1^{bO} \\ \mu_2^{bO} \end{bmatrix}, \begin{bmatrix} (\sigma_1^{bO})^2 + (\sigma^C_{(d,i),1})^2 + (\sigma^O_{(d,i),1})^2 & \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} \\ \rho^{bO}\sigma_1^{bO}\sigma_2^{bO} & (\sigma_2^{bO})^2 + (\sigma^C_{(d,i),2})^2 + (\sigma^O_{(d,i),2})^2 \end{bmatrix} \right). \tag{B.1}$$

Within eq. (B.1), the data for each drug-indication pair $(d,i) \in \mathcal{C}$ is known, while we seek to estimate bias distribution parameters μ_1^{bO}, μ_2^{bO}, σ_1^{bO}, σ_2^{bO}, and ρ^{bO}. Using eq. (B.1), define $LL_{(d,i)}(\mu_1^{bO}, \mu_2^{bO}, \sigma_1^{bO}, \sigma_2^{bO}, \rho^{bO})$ to be the log-likelihood of observing our results for drug-indication pair (d,i) under a specified set of bias distribution parameters. We proceed to select these parameters using maximum likelihood estimation, with a penalty for large values of $|\rho^{bO}|$. In particular, we obtain the parameters for our SOC pair in drug class $\mathcal{C}$ by solving the following optimization problem under regularization parameter $\lambda > 0$:

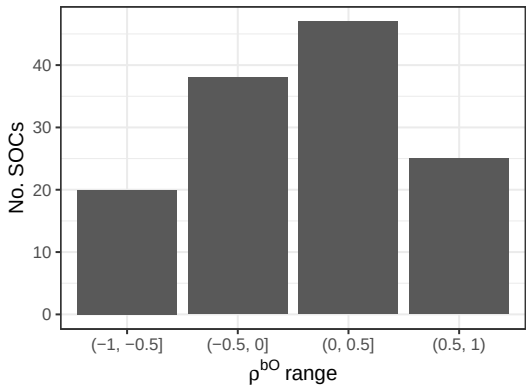
$$\max_{\mu_1^{bO}, \mu_2^{bO}, \sigma_1^{bO} \geq 0, \sigma_2^{bO} \geq 0, -1 < \rho^{bO} < 1} \frac{1}{|\mathcal{C}|} \sum_{(d,i) \in \mathcal{C}} LL_{(d,i)}(\mu_1^{bO}, \mu_2^{bO}, \sigma_1^{bO}, \sigma_2^{bO}, \rho^{bO}) - \lambda \cdot \log \frac{1 + |\rho^{bO}|}{1 - |\rho^{bO}|} \tag{B.2}$$

Note that the regularization term takes value 0 when $\rho^{bO} = 0$, and otherwise increases with $|\rho^{bO}|$.

To select the λ to use in our regularized model, we proceed by 10-fold cross-validation. In particular, we randomly partition our drug-indication pairs for our drug class. For each partition, we optimize the bias distribution parameters using eq. (B.2), fitted to all drug-indications except those in the selected partition; we separately estimate parameters for each λ in a set of candidate values $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.07, 0.1, 0.2, 0.3\}$. We then use those fitted parameters to obtain an out-of-sample log-likelihood for the drug-indication pairs within the selected partition. We repeat this process for each partition, obtaining a set of out-of-sample log-likelihoods for each drug-indication pair under each of our candidate λ values. We select the λ with the best out-of-sample performance and use that value to estimate the overall model using all the data from the drug class. The resulting estimates for our critical parameter ρ^{bO} are reported in Figure B3. Recalling the discussion in Section 3.1 of the mechanism behind the observational bias considered in this work, it is unsurprising that roughly half of the non-zero ρ^{bO} values are positive and the other half are negative—if a subgroup like diabetic patients have some side effects that constitute

a larger-than-average proportion of their reported ADRs (*e.g.*, renal and cardiovascular ADRs), then there must be a set of side effects that constitute a smaller-than-average proportion of their ADRs (*e.g.*, musculoskeletal and lymphatic ADRs).

Figure B3 ρ^{bO} **estimates**



Note. Each observation corresponds to a SOC paired with gastrointestinal disorders SOC. We estimate 130 values of ρ^{bO} across five drug classes.

B.3. Additional results

The final set of drug-PT pairs included in the analysis is reported in Table B1. Throughout this appendix section, we present a series of results comparing the proposed approach's performance to that of the two benchmark approaches.

Table B1 **Summary of drug-PT pairs included in the numerical analysis.**

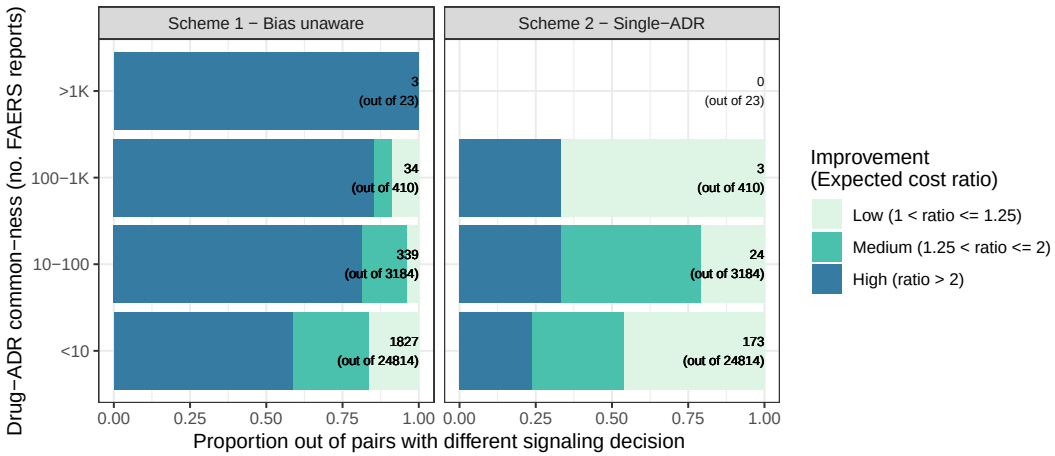
Drug Name	PTs	No. Signal (%)
Cyclophosphamide	2,560	930 (36%)
Rituximab	4,402	1,713 (39%)
Cetuximab	1,905	587 (31%)
Clopidogrel	1,973	553 (28%)
Cilostazol	1,349	561 (42%)
Dasatinib	1,949	427 (22%)
Sorafenib	2,650	967 (36%)
Chlorambucil	598	251 (42%)
Nilotinib	3,344	916 (27%)
Elotuzumab	520	224 (43%)
Vemurafenib	1,514	534 (35%)
Ruxolitinib	3,655	1,169 (32%)
Afatinib	1,310	461 (35%)
Obinutuzumab	702	272 (39%)
Total	28,431	9,565 (34%)

Note. Signals correspond to those made by our approach, assuming a z-score of 2 to calibrate M_{I}.

Performance by commonness of the ADR. Our approach reports benefits across all commonness buckets, as shown in Figure B4. Relative to Scheme 2, the cross-ADR improves about 0.7% of the pairs within each common-ness bucket except the largest ($> 1K$ reports), and about 7–13% when

compared to Scheme 1 across common-ness buckets. Of note, we find that the cross-ADR approach achieves superior performance for a large number of extremely rare ADRs (< 10 FAERS reports), which account for over 80% of all pairs with improvement. This is a key advantage of our approach; it has the potential to detect signals earlier in the post-marketing surveillance process (*i.e.*, with only a few reports).

Figure B4 Expected cost ratio by ADR commonness

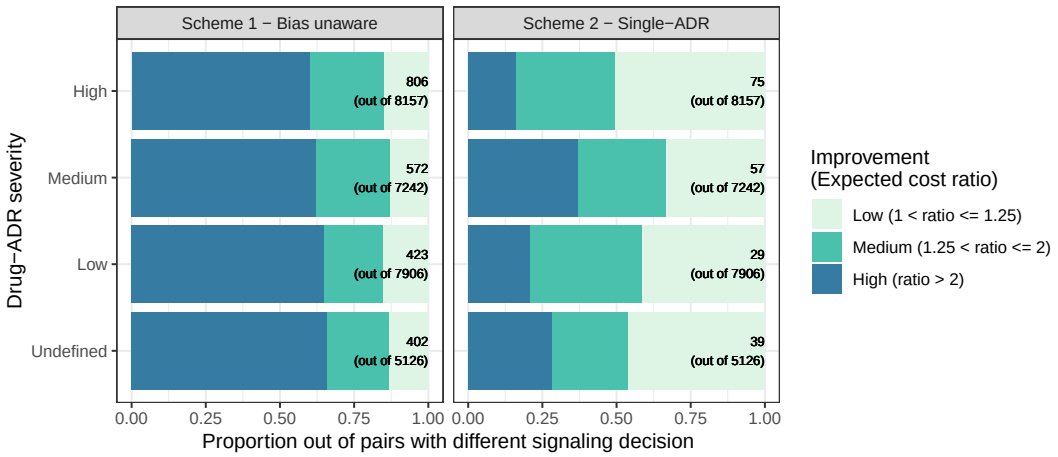


Note. The commonness bucket of a drug-ADR pair corresponds to the ADR raw count of FAERS reports. The number on the far right of each bucket corresponds to the number of drug-ADR pairs with different signaling decisions (compared to scheme 3), and the one in parentheses corresponds to the total number of drug-ADRs in that bucket.

Performance by ADR severity. Figure B5 shows the benefit of the cross-ADR approach in relation to Schemes 1 and 2, categorized by the severity of the ADR. We utilize the severity classification proposed by Lavertu et al. (2021). It is worth noting that the severity subgroups in our data tend to have a comparable number of pairs. Interestingly, the performance of the cross-ADR approach appears to be consistent across different ADR severities, as it shows comparable results for both high- and low-severity ADRs.

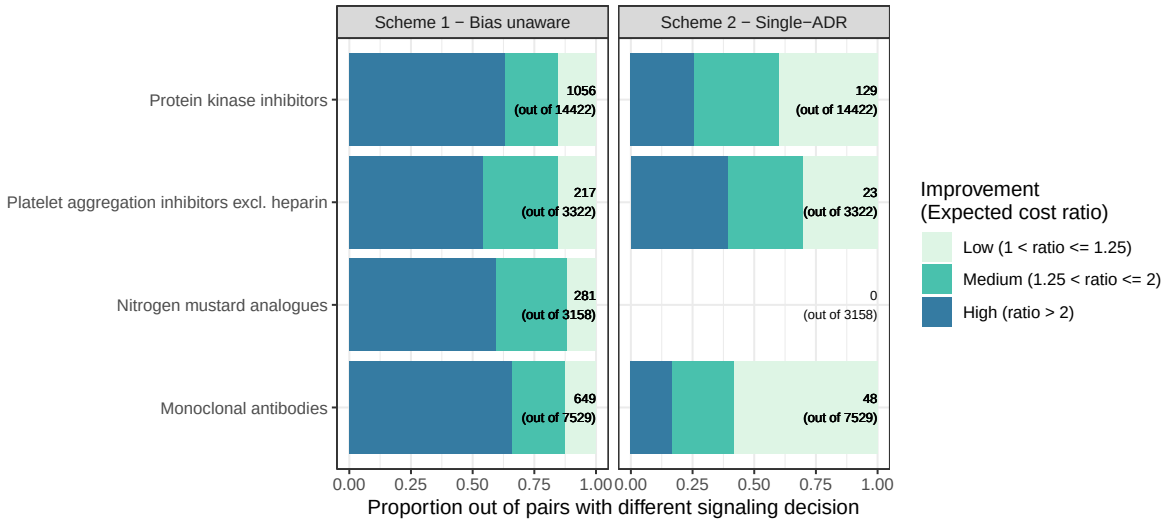
Performance by drug class. Figure B6 shows the improvement of the cross-ADR debiasing approach relative to Scheme 1 (left panel) and Scheme 2 (right panel) by drug class. The suboptimal signaling decisions made under the bias-unaware approach lead to expected costs that are twice as large as our approach for more than 50% of the pairs for all drug classes. Comparing the cross-ADR approach to Scheme 2, we observe comparable benefits across all but one drug class. In the case of nitrogen mustard analogues drug class, signaling differences are minimal due to the low correlation ρ^{bO}, which is near zero for all but three SOCs. Even in those SOCs (covering 715 PTs), the posterior mean difference between our method and the single-ADR approach remains low, explaining the alignment in signaling outcomes.

Figure B5 **Expected cost ratio by ADR severity**



Note. Severity classification is based on the PT of the pair. The number on the far right of each bucket corresponds to the number of drug-ADR pairs with different signaling decisions (compared to scheme 3), and the one in parentheses corresponds to the total number of drug-ADRs in that bucket.

Figure B6 **Expected cost ratio by drug class**



Note. The number on the far right of each drug class bucket corresponds to the number of drug-ADRs with different signaling decisions compared to scheme 3, and the one in parentheses corresponds to the total number of drug-ADR pairs in that bucket.

Appendix C: Model misspecification

In addition to the theoretical analysis of misspecification in the key correlation bias parameter ρ^{bO} (e-companion, Section E), we assess the impact of misspecification in all model parameters of the analytical model using estimates from the numerical analysis. Specifically, we focus on the same set of 28,431 drug-ADR pairs, and restrict to 26,307 pairs that are associated with SOCs other than

gastrointestinal disorders (our common ADR). Unlike our main numerical results, we compare the *expected cost at approval time* (noting that the ADRs have not been labeled at approval time since they have not been observed in clinical trials) (eq. (8)) rather than at signaling time to enable us to evaluate the impact of misspecification of all elements of our main model. While most parameter estimates are directly borrowed from Section 5.4, a few parameters are adjusted as follows:

- *Loss function:* In contrast to our main results that consider a smoothed lump-sum loss function, we consider a standard lump-sum loss function ($k = 0$) for computation tractability. We use the cost parameters M_{II}^S and M_{I}^S (calibrated for a z-score = 2) from our main results for each drug-PT pair.
- *Bias distribution:* We use the bias distribution parameters estimated at the drug-class level for each pair of common-rare SOCs.
- *Clinical trial distribution:* Unlike in the main analysis, we model clinical trial parameters for the rare ADR in the misspecification analysis, assuming a clinical trial effect size $\hat{\mu}_2^C = 0$ and estimating σ_2^C using a typical value of the clinical trial standard error at the SOC level (recall that the rare ADRs were not observed in the clinical trial for all our drug-ADR pairs, but similar ADRs within the same SOC have been observed).

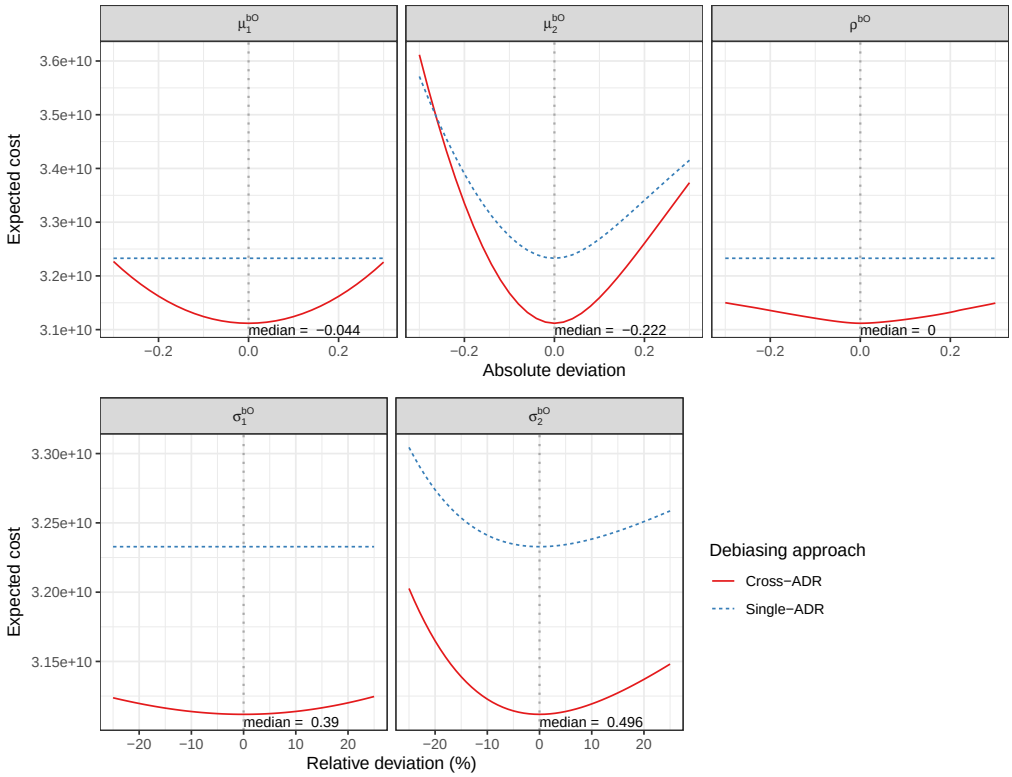
We vary the level of misspecification, depending on the parameter, within a moderate range of values. For positive-valued parameters (*e.g.*, σ_i^{bO}), we consider relative deviations: $\text{param}^{new} := \textit{deviation} \times \text{param}^{old}$. For all other parameters, we consider absolute deviations, *i.e.*, $\text{param}^{new} := \textit{deviation} + \text{param}^{old}$, with both positive and negative deviation depending on the parameter.[14] Note that if the parameter affects both the cross-ADR and the single-ADR benchmark, then both approaches make the signaling decision with the new, misspecified parameter. Note that all the parameters associated with the common ADR (*e.g.*, μ_1^{bO}) only affect our approach, not the single ADR benchmark.

C.1. Misspecification results

Our approach is generally robust to moderate misestimation.

Observational bias. Figure C7 shows that even with moderate misspecification, our approach consistently outperforms the single-ADR benchmark. Although the cost advantage decreases as misspecification increases, our method remains superior when the odds ratio parameter ($\exp(\mu_i^{bO})$) deviates within -26% to $+35\%$. In cases of severe underestimation of μ_2^{bO}, our method may perform worse. However, such large errors are unlikely because μ_2^{bO} is estimated from a large set of drug-indication pairs, reducing the risk of significant estimation error. Even with large deviations in the bias correlation parameter ρ^{bO} (up to ± 0.5), our approach still achieves a 1–2% reduction in expected costs. Our results also show strong robustness to estimation errors in the bias noise σ_i^{bO}. While errors in σ_2^{bO} affect both methods, our approach continues to maintain a performance advantage.

Figure C7 Misspecified observational bias parameters (median parameter across selected drug-PTs)



Generalization bias. Figure C8 illustrates the sensitivity of benefits to generalization bias parameter estimation errors. The correlation ρ^{bC} does not affect the performance under our selected subset of drug-ADRs (recall that we selected drug-ADRs with estimates $\sigma_1^{bC} = 0$). Similar to above, our approach performs better than the single-ADR approach for small-to-moderate estimation errors in μ_1^{bC}. The deviations in the rare ADR mean bias μ_2^{bC} affect both approaches in a similar way.

Sampling correlations. Figure C9 shows the sensitivity of the results to the sampling correlation parameters, ρ^{sC} and ρ^{sO}. Although we assume these parameters are zero, our approach continues to outperform the benchmark even with moderate deviations. The expected cost increases by less than 1.5% under the largest deviations considered. This is unsurprising, since we focus on drug-ADRs with well-estimated common ADRs.

Figure C8 Misspecified generalization bias parameters (median parameter across selected drug-PTs)

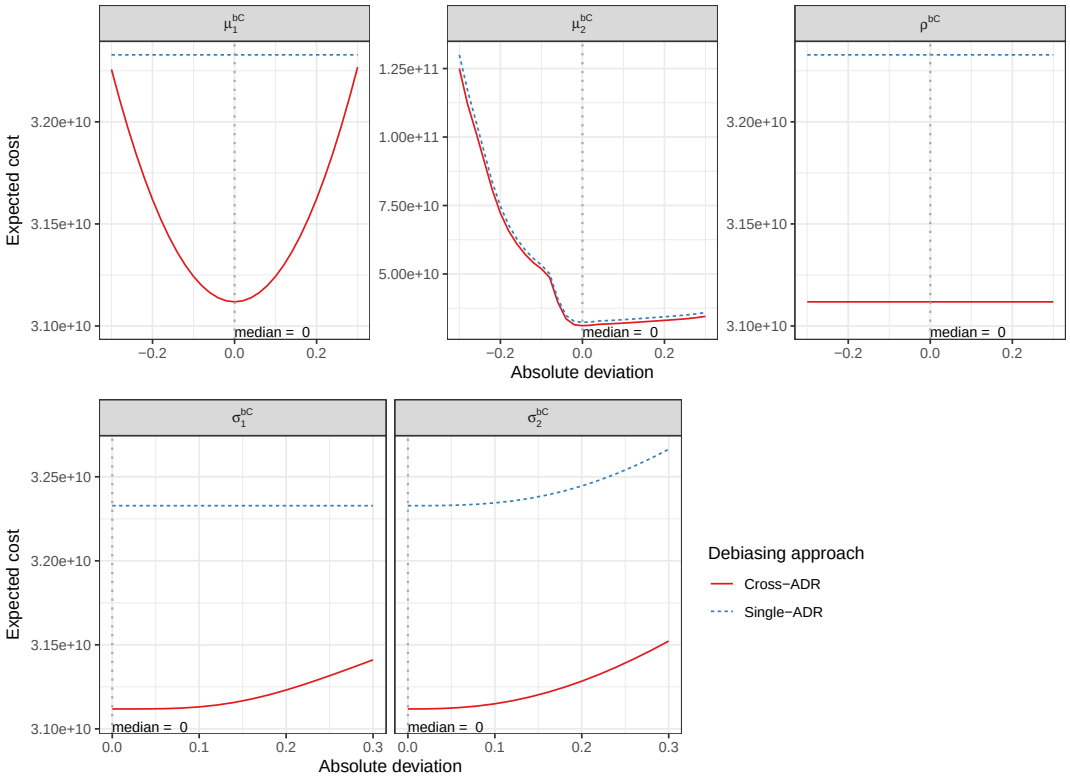
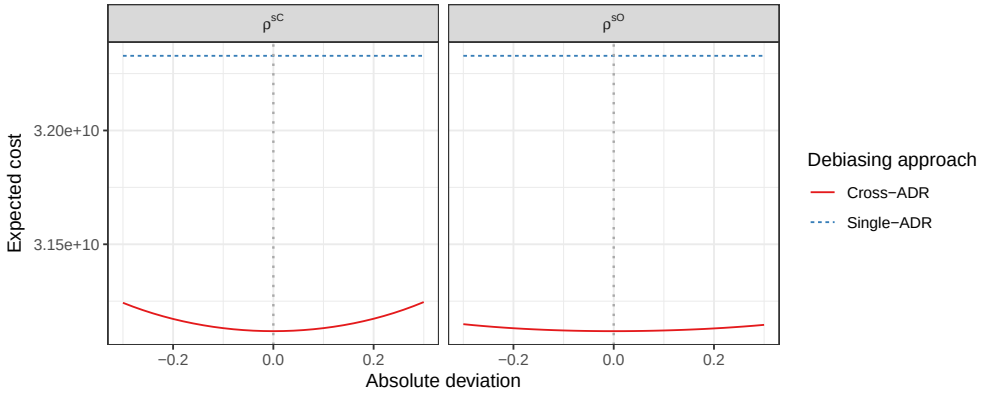


Figure C9 Misspecified sampling correlation parameters



Appendix D: Economic evaluation

D.1. Costs and loss functions

Define m_{ij} to be the expected number of future individuals taking drug i who will experience rare ADR j of interest. Some of these individuals would experience the ADR regardless of drug exposure; under a log risk ratio of $\beta \geq 0$ we would expect $e^{-\beta} m_{ij}$ individuals to experience the side effect without drug exposure, meaning $(1 - e^{-\beta}) m_{ij}$ cases could possibly be avoided by raising a safety signal. While the logic above is stated in terms of log risk ratios, recall that the log risk ratio and log odds ratio are approximately equal for a rare event. Defining $M_{\text{II}} := c_{ij} m_{ij}$, where c_{ij} is the economic cost of one case of our ADR j, we obtain the expected type II error cost by integrating over effect sizes $\beta \sim \mathcal{N}(\hat{\mu}_2^S, (\sigma_2^S)^2)$:

$$\mathcal{L}_S(\hat{\mu}_2^S, \sigma_2^S, 0) = M_{\text{II}} \int_0^\infty (1 - \exp(-x)) \frac{1}{\sigma_2^S} \phi\left(\frac{x - \hat{\mu}_2^S}{\sigma_2^S}\right) dx. \tag{D.3}$$

For symmetry, we consider a Type I error of the same functional form:

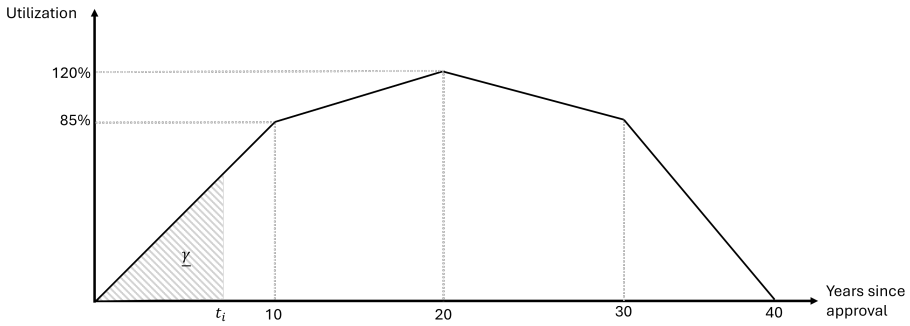
$$\mathcal{L}_S(\hat{\mu}_2^S, \sigma_2^S, 1) = M_{\text{I}} \int_{-\infty}^0 (1 - \exp(x)) \frac{1}{\sigma_2^S} \phi\left(\frac{x - \hat{\mu}_2^S}{\sigma_2^S}\right) dx. \tag{D.4}$$

One interpretation is that these represent smoothed versions of the lump-sum loss function ($k = 0$), capturing the fact that raising a safety signal may avert some fraction of, but not all, ADRs. The two key parameters to estimate are the costs of mistakenly signaling M_{I} and missing a signal M_{II}. We focus on estimating M_{II} and perform sensitivity analyses with respect to M_{I}.

The cost of missing a sign M_{II}—the harm cost—depends on the number of patients who would experience the adverse reaction (m_{ij}) and the corresponding cost of experiencing the ADR (c_{ij}). We estimate the expected number of patients who would experience the adverse reaction (if no signal is raised) by projecting the number of ADR cases that are yet to occur based on the age of the drug (*i.e.*, time since entering the market), the number of reports so far observed (adjusting for FAERS underreporting), and a standard drug utilization pattern. We quantify the cost of an ADR based on its severity.

A drug life-cycle can span over 30 years (Danzon and Kim 2002), so we assume that a molecule stays in the market for 40 years on average. Danzon and Kim (2002) and Duflos and Lichtenberg (2012) identify an inverted U-shape molecule utilization pattern with a market protection period between 8 and 12 years. Examples can be found in Figure 5 and Figure 6 of Duflos and Lichtenberg (2012) and Figure 2a of Danzon and Kim (2002). Therefore, we assume that the utilization of a molecule, including both brand-name and generic use, follows the pattern in Figure D10. The area under the curve, denoted by A, represents the theoretical cumulative utilization of the molecule

Figure D10 Lifetime of molecule utilization (brand-name + generic)



Note: The y-axis values 85% and 120% are relative to year 14 since market introduction, extrapolated from Duflos and Lichtenberg (2012) and Danzon and Kim (2002).

over its life cycle. Let us consider a drug with time since approval t. Then, given t, we define $\underline{\gamma}(t)$ as the theoretical cumulative utilization up to time t, namely,

$$\underline{\gamma}(t) = \begin{cases} \frac{0.85\times t^2}{20} & \text{if } t \in (0,10] \\ \frac{0.85\times 10}{2} + \frac{1}{2}(0.85+0.035t+0.5)\times(t-10) & \text{if } t \in (10,20] \\ \frac{0.85\times 10}{2} + \frac{1}{2}(0.85+1.2)\times 10 + \frac{1}{2}(1.2-0.035t+1.9)\times(t-20) & \text{if } t \in (20,30] \\ \frac{0.85\times 10}{2} + \frac{1}{2}(0.85+1.2)\times 20 + \frac{1}{2}(0.85-0.085t+3.4)\times(t-30) & \text{if } t \in (30,40]. \end{cases}$$

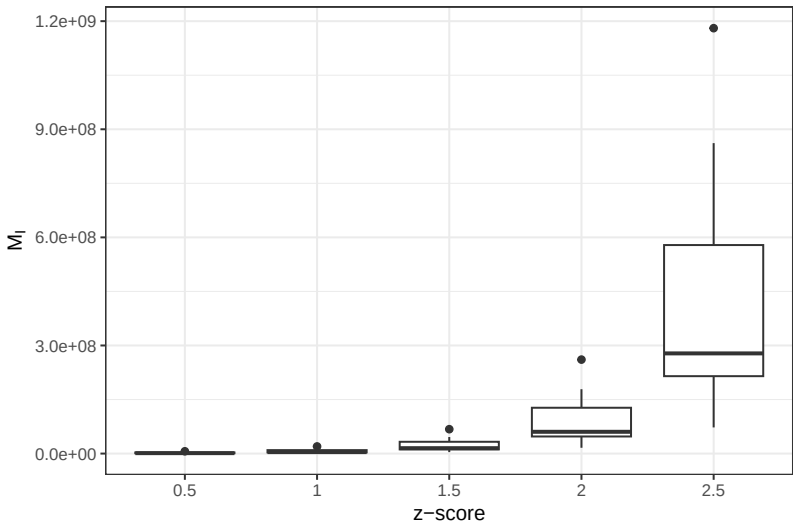
Similarly, we define $\overline{\gamma}(t) := A - \underline{\gamma}(t)$ as the theoretical cumulative utilization from time t until the end of the molecule lifetime. Then, for a given molecule or drug i, with time since approval t_i, we project the number of patients who would experience ADR j as follows:

$$m_{ij} = \frac{\overline{\gamma}(t_i)}{\underline{\gamma}(t_i)} \times \frac{r_{ij}}{6\%}, \tag{D.5}$$

where r_{ij} corresponds to the cumulative number of FAERS reports since approval, and the 6% adjustment captures the underreporting rate in FAERS (Hazell and Shakir (2006)). Note that r_{ij} and t_i can be computed from FAERS data. The fraction $\frac{\overline{\gamma}(t)}{\underline{\gamma}(t)}$ scales the future number of reports, so small fractions indicate that most reports have already been observed, while larger numbers indicate that most reports are yet to come, based on the drug's age and the theoretical utilization pattern in Figure D10. In this calculation, we assume the ADR reporting rate only varies based on the theoretical drug utilization rate. However, in practice, it has been reported that ADR reporting increases in the early years since approval (Hoffman et al. (2014)).

Estimating the costs of experiencing an ADR is difficult as it involves non-medical, indirect, and opportunity costs for patients and society. Since we study many ADRs, it is infeasible to estimate each cost separately. Instead, we extrapolate costs based on the severity of the ADR. We assign a cost of US$10,000, US$5,000, and US$1,000 depending on whether the ADR is high, medium, or low severity.[15] The following literature informs this cost range: Trumbo et al. (2021) reports per-patient costs of ADRs associated with iron products to about US$9,000 for high-severity ADRs and about

Figure D11 Summary of estimated drug-specific M_{I} for each z-score



Note. Each boxplot consists of one value for each unique drug.

US$2,000 for low-severity ones. Maity and Longo (2021) reports average costs (all ADRs) for TNF-α inhibitor biologics on the order of CAD$5,000, and Tissot et al. (2022) report costs on the order of €4,000 for ADRs that resulted in hospital admission. Pirmohamed et al. (2004) reports median ADR costs associated with length of stay, avoidability, and outcome on the order of US$3,000. The costs reported in these studies expand to non-medical, indirect, and opportunity costs. Thus, the cost of missing a signal is $M_{\text{II}} := m_{ij} \times c_{ij}$, where c_{ij} represents the drug i per-patient cost of experiencing ADR j.

To obtain the cost of false positive signals, M_{I}, we consider various z-score values $z \in \{0.5, \ldots, 2.5\}$ and solve for the value of $M_{\text{I}}(z)$ that makes the expected loss of signaling and not signaling equal under the specific z (*i.e.*, apply the optimality condition in Lemma 1). To minimize the impact of outliers, in our final evaluation, we assume that the cost of raising a false signal for a given drug does not vary based on the specific ADR for which we are raising the signal. Basically, raising a false signal can be more costly on a popular drug versus an unpopular one, but it does not depend on what we signal for. Then, for each drug, we use the median value, denoted by $\bar{M}_{\text{I}}(z)$, across all ADRs. For a z-score of 2, $\bar{M}_{\text{I}}(z) \in [1.6\text{e}{+07},\ 2.6\text{e}{+08}]$ with a median of 6.0e+07 across all PTs. Figure D11 summarizes the estimated M_{I} values for each z-score.

D.2. Computing r_{ij} and t_i from FAERS data

The cumulative number of FAERS reports r_{ij} for each drug-ADR pair can be obtained directly from FAERS data by counting all reports up to 2020. The time since approval t_i can be inferred as follows: for each drug i, we identify the year of the first-percentile report as the year of the drug's approval and measure the drug's age as of 2020, *i.e.*, $t_i := 2020 - \text{approval year}$.

D.3. Limitations

Our procedure has several limitations. First, we estimate the cost of missing a safety signal, M_{II}, based on ADR severity, associated costs, and projected future cases. This projection relies on factors such as drug age, market lifespan (assumed to be 40 years), cumulative FAERS cases, and an assumed 6% FAERS reporting rate. However, this approach may inflate projections for certain drugs, particularly those recently approved or receiving heightened public attention due to lawsuits.

Additionally, our projections assume a 40-year drug lifespan, which may overstate M_{II}. To address this, we examine shorter lookahead periods (5 and 10 years), finding that while absolute cost estimates change, the relative performance of our approach remains consistent (results not shown).

Appendix References

Danzon PM, Kim JD (2002) The life cycle of pharmaceuticals: A cross-national perspective. Technical report, Office of Health Economics, London UK.

Duflos G, Lichtenberg FR (2012) Does competition stimulate drug utilization? the impact of changes in market structure on us drug prices, marketing and utilization. *International Review of Law and Economics* 32(1):95–109.

Hazell L, Shakir SA (2006) Under-reporting of adverse drug reactions. *Drug safety* 29(5):385–396.

Hoffman KB, Dimbil M, Erdman CB, Tatonetti NP, Overstreet BM (2014) The weber effect and the united states food and drug administration's adverse event reporting system (faers): analysis of sixty-two drugs approved from 2006 to 2010. *Drug safety* 37:283–294.

Maity T, Longo C (2021) Pragmatic pharmacoeconomic analyses by using post-market adverse drug reaction reports: an illustration using infliximab, adalimumab, and the canada vigilance adverse reaction database. *BMC Health Services Research* 21:1–12.

Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK, Breckenridge AM (2004) Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj* 329(7456):15–19.

Tissot M, Valnet-Rabier MB, Stalder T, Limat S, Davani S, Nerich V (2022) Epidemiology and economic burden of "serious" adverse drug reactions: Real-world evidence research based on pharmacovigilance data. *Therapies* 77(3):291–300.

Trumbo H, Kaluza K, Numan S, Goodnough LT (2021) Frequency and associated costs of anaphylaxis-and hypersensitivity-related adverse events for intravenous iron products in the usa: an analysis using the us food and drug administration adverse event reporting system. *Drug safety* 44:107–119.

that debias the DA estimate of a drug's effect on a rare ADR using the DA estimate for a more common ADR, as we propose. Our strategy could enhance any existing DA methods in practice.

There is substantial literature on using longitudinal data sources for postmarketing surveillance, such as electronic health record (EHR) data and medical claims data. These data are richer than spontaneous reporting databases, containing detailed patient covariates (EHR data) and longitudinal records of drug exposures and adverse events. Leveraging this richness, numerous methods incorporating temporal trends and time-varying covariates have been developed to investigate safety signals (Arnaud et al. 2017), with evidence from the U.S. (Ryan et al. 2013) and Europe (Schuemie et al. 2013) showing strong performance. In the operations literature, Ahuja et al. (2021) use logistic regression and matching designs to assess drug–ADR associations, demonstrating that the inaccurate black-box warning on rosiglitazone could have been avoided. Since such methods rely on careful cohort construction, covariate selection, and matching, they are mainly used to confirm rather than automatically generate new signals; for instance, the FDA has used EHR data accessed via its Sentinel system this way. Our work complements this stream by focusing on generating higher-quality signals for later evaluation with longitudinal databases.

Though much of the literature for longitudinal databases of health records relates to confirming safety signals, there is also significant literature on generating new safety signals (our task of "hypothesis generation") from a variety of observational data sources, including EHR data. Sequential statistical testing (SST) can be used to monitor ADRs using longitudinal health record data, with classical approaches such as the sequential probability ratio test (SPRT) (Wald 1945) and maximized SPRT (Brown et al. 2007). An exemplary example of this type of approach within the operations literature can be found in Goh et al. (2015), where the authors propose an algorithm that can be used for continuous monitoring of multiple potentially correlated ADRs; their approach provides Type I error control, is robust to distributional assumptions, and takes into account the temporal sequence of using the drug and experiencing the ADR. Data sources beyond spontaneous reporting databases or longitudinal health records have also been considered to generate new safety signals—these include molecular fingerprints (Vilar et al. 2011), medical literature findings (Xu and Wang 2014, Shetty and Dalal 2011), and social media data (Liu and Chen 2015, Yeleswarapu et al. 2014, Li et al. 2020). In this literature, authors often propose to obtain signals from multiple data sources and combine them to obtain more confidence before raising a final safety signal (Li et al. 2020, Harpaz et al. 2013). Given its scale, spontaneous reporting data is likely to remain an important data source for safety signaling, either as a standalone source or in combination with others, motivating our focus on improving signaling quality by removing bias.

Methodologically, our work draws on causal inference research combining observational and experimental data, including reweighting RCT data (*e.g.*, Hartman et al. (2015)), extrapolating