arXiv:2404.14213v1 [stat.ME] 22 Apr 2024

An Exposure Model Framework for Signal Detection based on Electronic Healthcare Data

Louis Dijkstra[a], Tania Schink[a], and Ronja Foraita[*a]

[a]Leibniz Institute for Prevention Research & Epidemiology – BIPS,
Achterstraße 30, 28359 Bremen, Germany

April 23, 2024

Abstract

Despite extensive safety assessments of drugs prior to their introduction to the market, certain adverse drug reactions (ADRs) remain undetected. The primary objective of pharmacovigilance is to identify these ADRs (i.e., signals). In addition to traditional spontaneous reporting systems (SRSs), electronic health (EHC) data is being used for signal detection as well. Unlike SRS, EHC data is longitudinal and thus requires assumptions about the patient's drug exposure history and its impact on ADR occurrences over time, which many current methods do implicitly.

We propose an exposure model framework that explicitly models the longitudinal relationship between the drug and the ADR. By considering multiple such models simultaneously, we can detect signals that might be missed by other approaches. The parameters of these models are estimated using maximum likelihood, and the Bayesian Information Criterion (BIC) is employed to select the most suitable model. Since BIC is connected to the posterior distribution, it servers the dual purpose of identifying the best-fitting model and determining the presence of a signal by evaluating the posterior probability of the null model.

We evaluate the effectiveness of this framework through a simulation study, for which we develop an EHC data simulator. Additionally, we conduct a case study applying our approach to four drug-ADR pairs using an EHC dataset comprising over 1.2 million insured individuals. Both the method and the EHC data simulator code are publicly accessible as part of the `R` package `https://github.com/bips-hb/expard`.

Keywords: electronic healthcare data; exposure model; pharmacovigilance; simulator; spontaneous reporting systems

*Corresponding author. E-mail: `foraita@leibniz-bips.de`

1 Introduction

Despite the thorough examination of drugs for potential side effects before their market release, some adverse drug reactions (ADRs) may go unnoticed until after the drug enters the market [60, 47, 22, 6, 58]. This could be due to several reasons. The pivotal randomized clinical trials (RCTs) are designed with a focus on assessing efficacy, resulting in sample sizes often inadequate for detecting rare safety outcomes [20, 45]. In these trials, patients typically need to meet multiple inclusion criteria, and vulnerable groups, such as pregnant women, the elderly [8], and individuals with multiple health conditions, are frequently either excluded or underrepresented [31, 17]. Furthermore, patients in RCTs are monitored for a limited duration only, making it difficult to uncover any potential long-term effects. For that reason, pharmacovigilance [60, 58, 18] plays a pivotal role in ensuring the safety of pharmaceutical products.

With the aim of promptly identifying drugs that may pose health risks, spontaneous reporting systems (SRSs) have been established over the years [39]. Healthcare professionals, pharmaceutical companies, and, in some cases, patients [29] can submit a spontaneous report to such a system when they suspect a drug may be associated with a previously unknown reaction [47, 6, 58, 41]. These reports are collected, cleaned, stored, and subsequently analyzed by a committee of medical experts [4, 36, 3]. Due to the sheer volume of accumulated reports typically contained in these systems [1], the idea was to present the committee of medical experts with an automatically curated list of drug-ADR pairs that needed their attention, rather than the raw reports themselves [50, 51, 2]. Each entry on such a list is referred to as a a *signal*. Therefore, the process of creating such a shortlist is also known as *signal detection* [58].

Even though SRSs form the cornerstone of pharmacovigilance, they also come with several limitations. First, the total number of patients exposed to the drug or experiencing the ADR is essentially unknown. Reports are submitted only when a patient has both been exposed to the drug *and* has experienced the ADR, leading to what is known as the unknown denominator problem [16]. Secondly, there is potentially an over- or underreporting bias [41, 51, 30, 53, 24]. Newly introduced drugs, for example, may attract more attention from healthcare professionals and are, thus, more likely to be reported. Third, the decision on what information to include in the report is made on an individual basis. For instance, one might choose to exclude a drug if it is deemed too unlikely to have caused the ADR or if the patient's exposure is considered too far in the past.

Over the last two decades, alongside SRS data, pharmacovigilance has begun considering longitudinal data as well, specifically *electronic healthcare* (EHC) *data* [10, 44, 12, 26, 32, 42, 48]. EHC data include individual patients' drug prescriptions and medical events over time, along with personal details such as age, sex, and place of residence [47, 9, 62]. This

kind of data, to some extent, mitigates the limitations of SRS data mentioned earlier [38]. The total number of patients exposed to the drug *and/or* experiencing the ADR is known; there is a reduction in over- and underreporting bias [21], and it eliminates the need to make choices about what to report. An additional advantage is that EHC data sets can be very extensive, containing up to millions of patients [23].

Another distinction between EHC and SRS data lies in their timeliness [52]. Reports are typically submitted promptly to SRSs, whereas EHC data often experience delays. Consequently, SRS data play a crucial role in detecting ADRs of drugs recently introduced to the market. On the other hand, the richness of EHC data may prove beneficial in identifying more subtle associations between the drug and ADRs, a task that might be challenging on the basis of SRS data alone [49].

A plethora of signal detection methods have been proposed for EHC data [47, 42, 9, 62, 54, 55, 27]. To address its longitudinal nature, many of these methods convert EHC data into a format resembling SRS data and utilize techniques initially developed for SRSs [62]. Other approaches such as LASSO [13] and Random Forests (RFs) condense a patient's exposures and ADR occurrences over time into a limited number of variables while striving to retain some of its temporal information [13].

By transforming EHC data in this way, one implicitly makes assumptions about the temporal relationship between drug exposure and the ADR [35]. Van Gaalen et al. [54, 55] introduce the concept of an *exposure model*, i.e., a formal description of how exposure to the drug affects the risk for a patient to experience the ADR over time. The idea is to define multiple exposure models, each attempting to capture a different type of temporal relationship. For instance, one can establish an exposure model for withdrawal effects, where the risk of experiencing the ADR peaks just after the patient's exposure stops and diminishes rather quickly. Alternatively, a model can be defined where the effect of the exposure is long-term, and the risk increases gradually over time [54]. By considering multiple of these models simultaneously for each drug-ADR pair, the hope is to be able to identify many different temporal relationships, some of which may not be discernible using conventional approaches since the effect is diluted. Van Gaalen et al. [54, 55] define several such exposure models. However, the range of models is limited and it is unclear how the parameters of these models are to be estimated on the basis of the data. Additionally, their approach focuses on a single drug-ADR pair, leaving ambiguity about its applicability in a pharmacovigilance context where multiple drug-ADR pairs are considered simultaneously.

An alternative approach involves employing cubic B-splines to characterize the connection between the drug-ADR pair [27]. While this approach offers flexibility, it comes with drawbacks, as it does not explicitly define the nature of the relationship between the drug and ADR pair. More critically, the application of cubic B-splines in the traditional pharma-

covigilance setting, where multiple drug-ADR pairs are simultaneously considered, is unclear as well [27].

In this work, we present a comprehensive exposure model framework for pharmacovigilance based on EHC data to address some of the aforementioned limitations. We formally define the concept of an exposure model. In addition, we propose eight exposure models, known to occur, at least approximately, in real-world data. It is important to note that these serve as examples, and any other exposure model can be defined within the framework. The exposure models' parameters are then estimated using maximum likelihood.

We advocate the use of the Bayesian Information Criterion (BIC) for model selection, given its focus on both the quality of data fit (expressed by the model's likelihood) and the model's complexity (quantified in terms of the number of parameters). An additional advantage of the BIC is its close relationship to the posterior probability of the exposure model. By utilizing the posterior probability of the null model, indicating no association between the drug and the ADR (see Section 2.5), our approach can effectively achieve two objectives: 1) determining the presence of an association, and 2) identifying which exposure model among those considered best fits the data.

The posterior probabilities of the exposure models offer an added advantage, making the approach well-suited for pharmacovigilance by providing a basis for which drug-ADR pairs are to be considered a signal. The weakest signal shows the pair with the highest probability for the null model indicating no association, while the pair with the lowest posterior probability is considered the strongest signal.

To assess the effectiveness of our approach, we perform a simulation study. We offer a simulator for EHC data, which enables users to utilize the exposure model of their choice to simulate the occurrences of ADRs over time. We then assess the performance of the method in terms of its capability to both detect an association and accurately identify the correct exposure model.

Furthermore, we demonstrate the applicability of our approach through a case study based on data from the German Pharmacoepidemiological Research Database[23] (GePaRD), see Section 6. In this case study, we examine four drug-ADR pairs, three of which are known to be associated, and one serving as a negative control. The literature provides information on the longitudinal nature of the associations for these three positive pairs. We assess the extent to which our method can accurately identify the correct exposure model.

The paper is structured as follows: We begin by formally defining an EHC dataset in Section 2.1. Subsequently, in Section 2.2, we introduce a formal definition of an exposure model. Section 2.3 contains eight examples of such models. We then detail how the parameters of the exposure models can be estimated using maximum likelihood, with analytical solutions derived for three of the eight models. The remaining models are solved numerically.

Section 2.5 considers model selection and the decision criterion for determining whether a drug is associated with an ADR.

Subsequently, we introduce the simulator for EHC data in Section 3 and outline the simulation set-up for the simulation study in Section 4. The method for assessing the performance in the simulation study is described in Section 5. Following this, we introduce our case study in Section 6. The outcomes of both the simulation and the case study are presented in Section 7. We conclude with some final remarks and discussion in Section 8.

All the code is publicly available online. Both the simulator and the implementation of the method are available in form of the `R` package `expard` at `https://github.com/bips-hb/expard`. The code related to the simulation study and case study can be found at `https://github.com/bips-hb/expard-simulation-study` and `https://github.com/bips-hb/expard-case-study`, respectively.

2 Methods

2.1 A Formalization of Electronic Healthcare Data

Electronic healthcare data contains for multiple patients 1) the drugs they were exposed to, and 2) the ADRs they experienced over time. We denote the number of drugs on the market by m; the number of registered ADRs and the number of observed patients are denoted by n and N, respectively. The number of time points for which a patient was observed, can differ from patient to patient. We denote the total number of time points for the k-th patient by $T_k \geq 1$. We assume that each patient was observed continuously, i.e., without interruptions.

We represent the k-th patient's drug exposure to the i-th drug over time as a random binary T_k-dimensional vector:

$$\boldsymbol{X}_i^k = \left(X_i^k(1), X_i^k(2), \ldots, X_i^k(T_k)\right),$$

where $X_i^k(t) = 1$ if the k-th patient was exposed to the i-th drug at time point t, and 0 otherwise. Likewise, the occurrences of the j-th ADR are represented by the random T_k-dimensional binary vector

$$\boldsymbol{Y}_j^k = \left(Y_j^k(1), Y_j^k(2), \ldots, Y_j^k(T_k)\right),$$

where $Y_j^k(t) = 1$ if the k-th patient had the j-th ADR at time point t, and 0 otherwise. Since there are m drugs, we can represent all drug exposures for a patient k as a set of m different T_k-dimensional binary vectors, i.e.,

$$\boldsymbol{P}_{\text{drugs}}^k = \left\{\boldsymbol{X}_1^k, \boldsymbol{X}_2^k, \ldots, \boldsymbol{X}_m^k\right\}.$$

Likewise, we can represent the ADR history for patient k as a set with n binary vectors:

$$\boldsymbol{P}^k_{\text{ADRs}} = \left\{\boldsymbol{Y}^k_1, \boldsymbol{Y}^k_2, \ldots, \boldsymbol{Y}^k_n\right\}.$$

The k-th patient is represented by both his/her drug exposure and ADR history, i.e.,

$$\boldsymbol{P}_k = \left\{\boldsymbol{P}^k_{\text{drugs}}, \boldsymbol{P}^k_{\text{ADRs}}\right\}.$$

And lastly, an EHC data set is then a collection of N patients:

$$\mathbf{EHC} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_N\}.$$

Observations are denoted by lower-case letters: $\mathbf{ehc} = \{\boldsymbol{p}_k\}_{k=1}^N$ is a given EHC data set, where $\boldsymbol{p}_k = \{\boldsymbol{p}^k_{\text{drugs}}, \boldsymbol{p}^k_{\text{ADRs}}\}$ is the k-th patient. The set $\boldsymbol{p}^k_{\text{drugs}}$ represents the observed drug exposures: $\boldsymbol{x}^k_i = (x^k_i(1), x^k_i(2), \ldots, x^k_i(T))$ for $i = 1, 2, \ldots, m$. Similarly, the set $\boldsymbol{p}^k_{\text{ADRs}}$ represents the observed ADRs: $\boldsymbol{y}^k_j = (y^k_j(1), y^k_j(2), \ldots, y^k_j(T))$ for $j = 1, 2, \ldots, n$.

One commonly assumes *patient independence*, meaning that the joint probability density function of **EHC** can be factorized as

$$\mathbb{P}(\mathbf{EHC} = \mathbf{ehc}) = \prod_{k=1}^{N} \mathbb{P}(\boldsymbol{P} = \boldsymbol{p}_k),$$

where $\mathbb{P}(\boldsymbol{P})$ denotes the probability density function of a single patient. For readability's sake, we assume in the following that all patients have been observed for the same number of time points, i.e., $T = T_1 = T_2 = \ldots = T_N$. It is straightforward to extend these signal detection methods to deal with varying observation times. Throughout this paper, we mainly consider single drug-ADR pairs. We, therefore, omit the subscripts i and j wherever possible. For ease of notation, we write

$$\boldsymbol{X}(1:t) = (X(1), X(2), \ldots, X(t-1), X(t)) \in \{0,1\}^t$$

for the drug exposures for that patient from time point 1 to t.

2.2 The Exposure Model

In pharmacovigilance, we are commonly interested in the joint probability distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ for all drug-ADR pairs in the data set. Specifically, we are interested in whether they are independent, i.e.,

$$\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}\right) = \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right)\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right)$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^T$. A full specification of the joint probability density function of $(\boldsymbol{X}, \boldsymbol{Y})$ is infeasible, since it requires to specify a total of 2^{2T} probabilities. The model, therefore,

has to be simplified. As mentioned in Section 1, signal detection methods that have been proposed in the past do this implicitly, most commonly by transforming the EHC data for a given drug-ADR pair to a single 2×2 contingency table. There are a variety of ways to do this, see, for example, the work by Zorych et al. [62].

Instead of the full probability distribution, we consider here the conditional distribution of the ADR history given the exposures to the drug over time:

$$\mathbb{P}\left(\boldsymbol{Y} \mid \boldsymbol{X}\right) = \prod_{t=1}^{T} \mathbb{P}\left(Y(t) \mid \boldsymbol{X}(1:t)\right). \tag{1}$$

Modelling the full probability distribution $\mathbb{P}\left(\boldsymbol{X}, \boldsymbol{Y}\right) = \mathbb{P}\left(\boldsymbol{Y} \mid \boldsymbol{X}\right)\mathbb{P}(\boldsymbol{X})$ would require to model the drug exposure to the drug over time, i.e., $\mathbb{P}(\boldsymbol{X})$, which is not of direct interest. We assume that the occurrences of the ADR at different time points are independent given the drug exposure history. This can be a rather strong assumption for certain types of ADRs, e.g., anaphylaxis and myocardial infarction. We address this in the discussion, see Section 8. Note that we can express $\mathbb{P}\left(\boldsymbol{Y}(t) \mid \boldsymbol{X}\right)$ as $\mathbb{P}\left(\boldsymbol{Y}(t) \mid \boldsymbol{X}(1:t)\right)$ since the occurrence of the j-th ADR at time point t is independent of drug exposures in the future given the drug exposures up to that point in time.

Here, we set out to create a framework for modelling the conditional probability distribution from eq. (1). To this end, we first define a *risk level* as a value between $[0, 1]$, where 0 represents that the patient is, at that point in time, at 'minimal' risk of experiencing the ADR and 1 represents 'maximal' risk (its precise meaning becomes clear later). In addition, we define $\mathcal{D}_T = \{\{0,1\}^t : t = 1, 2, \ldots, T\}$ be the set of all binary vectors of length $t = 1, 2, \ldots, T$. It, thus, represents all possible drug exposure histories. We propose to represent the conditional distribution in terms of an exposure model:

Definition 1 *An exposure model $\mathcal{M}$ for a drug-ADR pair is given by the tuple*

$$\langle \pi_1, \pi_0, T, r_{\mathcal{M}}(\cdot; \boldsymbol{\theta}), \Theta_{\mathcal{M}} \rangle, \tag{2}$$

*where $\pi_1, \pi_0 \in [0, 1]$, T is the number of time points, and $r_{\mathcal{M}} : \mathcal{D}_T \to [0, 1]$ is the **risk function** that maps each binary vector in $\mathcal{D}_T$ to a risk level in the interval $[0, 1]$. The risk function is parameterized by the m-dimensional parameter vector $\boldsymbol{\theta} = (\boldsymbol{\xi}; \boldsymbol{\phi}) \in \Theta_{\mathcal{M}}$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_q) \in \Xi_{\mathcal{M}}$ are continuous and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s) \in \Phi_{\mathcal{M}}$ are discrete parameters $(m = q + s \geq 0)$.*

Utilizing the exposure model $\mathcal{M}$, we can express the conditional probability of the ADR occurring at time point t, given the drug history up to that point, as

$$\mathbb{P}_{\mathcal{M}}\left(Y(t) \mid \boldsymbol{X}(1:t)\right) = (\pi_1 - \pi_0)\, r_{\mathcal{M}}(\boldsymbol{X}(1:t); \boldsymbol{\theta}) + \pi_0.$$

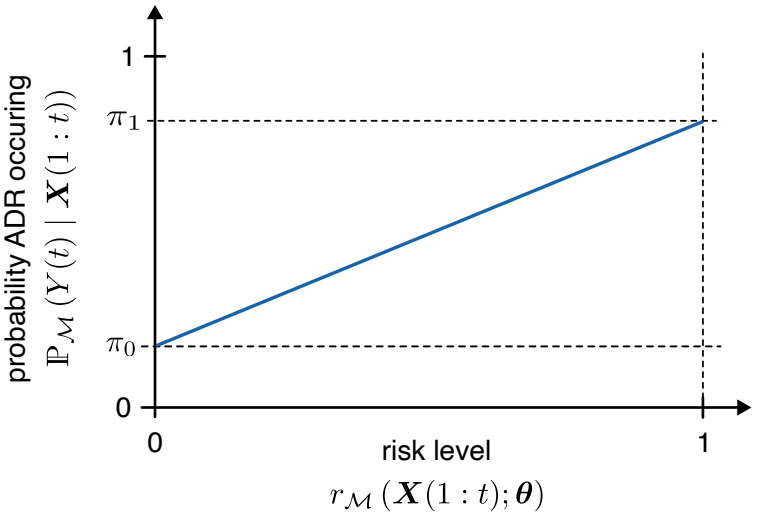

Figure 1: The probability of the ADR occurring at time point t, denoted as $\mathbb{P}_{\mathcal{M}}(Y(t) \mid \boldsymbol{X}(1:t))$, plotted against the risk level represented by $r_{\mathcal{M}}(\boldsymbol{X}(1:t);\boldsymbol{\theta})$. At the minimal risk level (0), the probability of the ADR taking place is π_0. Conversely, at the maximal risk level (1), the probability is π_1.

In other words, the probability is π_0 when the patient is at minimal risk ($r_{\mathcal{M}}(\boldsymbol{X}(1:t);\boldsymbol{\theta}) = 0$) and π_1 when the patient is at maximal risk ($r_{\mathcal{M}}(\boldsymbol{X}(1:t);\boldsymbol{\theta}) = 1$). This relationship is visualized in Figure 1, where the x-axis represents the risk level and the y-axis signifies the probability of the ADR taking place. As the risk level varies, this probability can assume any value within the interval $[\pi_0, \pi_1]$. Consequently, the conditional probability from equation (1) can be expressed as the product

$$\mathbb{P}_{\mathcal{M}}(\boldsymbol{Y} \mid \boldsymbol{X}) = \prod_{t=1}^{T} \mathbb{P}_{\mathcal{M}}(Y(t) \mid \boldsymbol{X}(1:t)) = \prod_{t=1}^{T} \left[(\pi_1 - \pi_0)\, r_{\mathcal{M}}(\boldsymbol{X}(1:t);\boldsymbol{\theta}) + \pi_0\right].$$

In the next section, we propose eight exposure models that mimic various known relationships between drugs and ADRs. We provide examples of drug-ADR pairs that are known to approximately follow those models. In Section 2.4 we discuss how to fit an exposure model $\mathcal{M}$ to a specific data set. It is important to note that the concept of an exposure model is of course not limited to the eight examples presented here; one can define others as well.

2.3 Examples of Exposure Models

Here, we propose eight exposure models by specifying the risk function $r_{\mathcal{M}}$ and the associated parameter space $\Theta_{\mathcal{M}}$ for each model.

No Assocation ($\mathcal{M}_0$)

The no association exposure model reflects the case when there is no association between the drug and ADR in question, i.e., $\mathbb{P}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{P}(\boldsymbol{X})\mathbb{P}(\boldsymbol{Y})$. In terms of the risk function, we can represent this case as

$$r_{\mathcal{M}_0}\left(\boldsymbol{X}(1:t)\right) = 0 \qquad \text{for all } \boldsymbol{X} \in \mathcal{D}_T.$$

The parameter space $\Theta_{\mathcal{M}_0}$ is the empty set $\emptyset$. In other words, the patient is at 'minimal risk' independent of his/her exposure to the drug. See, for example, Figure 2a. The drug exposure over time is represented by the x-axis, where the exposed period from $t = 5$ to $t = 10$ is denoted by the shaded area. The y-axis is the risk level over time given the exposure. As you can see, the risk level is constant and zero throughout.

Current Use ($\mathcal{M}_{\text{current use}}$)

The current use exposure model represents the case where the patient's risk level is elevated when the patient is exposed and returns to zero the moment he/she ceases to be exposed. Formally, we can express this as

$$r_{\mathcal{M}_{\text{current use}}}\left(\boldsymbol{X}(1:t)\right) = X(t) \qquad \text{for all } \boldsymbol{X} \in \mathcal{D}_T.$$

The parameter space is $\Theta_{\mathcal{M}_{\text{current use}}} = \emptyset$ as well. See Figure 2b for an example. One can see that the risk level is maximal ($= 1$) during exposure and minimal ($= 0$) when not exposed. An examples of a drug-ADR pair that, at least approximately, follow this pattern are oral corticosteroids and fractures [57].

Withdrawal Effects ($\mathcal{M}_{\text{withdrawal}}$)

The risk of experiencing withdrawal effects is highest quickly after a patient is no longer exposed and decreases with time. We can model this with the following risk function where, for all $\boldsymbol{X} \in \mathcal{D}_T$,

$$r_{\mathcal{M}_{\text{withdrawal}}}\left(\boldsymbol{X}(1:t); \rho\right) = \begin{cases} 0 & \text{if never/currently exposed, and} \\ \exp\left(\rho \cdot f_{\text{last}}(\boldsymbol{X}(1:t))\right) & \text{otherwise,} \end{cases}$$

where $\rho \in \Theta_{\mathcal{M}_{\text{withdrawal}}} = \mathbb{R}_+$ is the risk function's only parameter denoting the rate with which the risk level decreases, and the function $f_{\text{last}}(\cdot)$ returns the number of time points since the patient's last exposure, i.e.,

$$f_{\text{last}}\left(\boldsymbol{X}(1:t)\right) = t - \max\left\{\tau \in \{1, 2, \ldots, t-1\} \text{ such that } X(\tau) = 1\right\}. \tag{3}$$

Figure 2c and 2d show two examples where the rate parameter ρ is either 1 or $\frac{1}{2}$. A well-known example of drugs that elicit such a response are opioids [28].

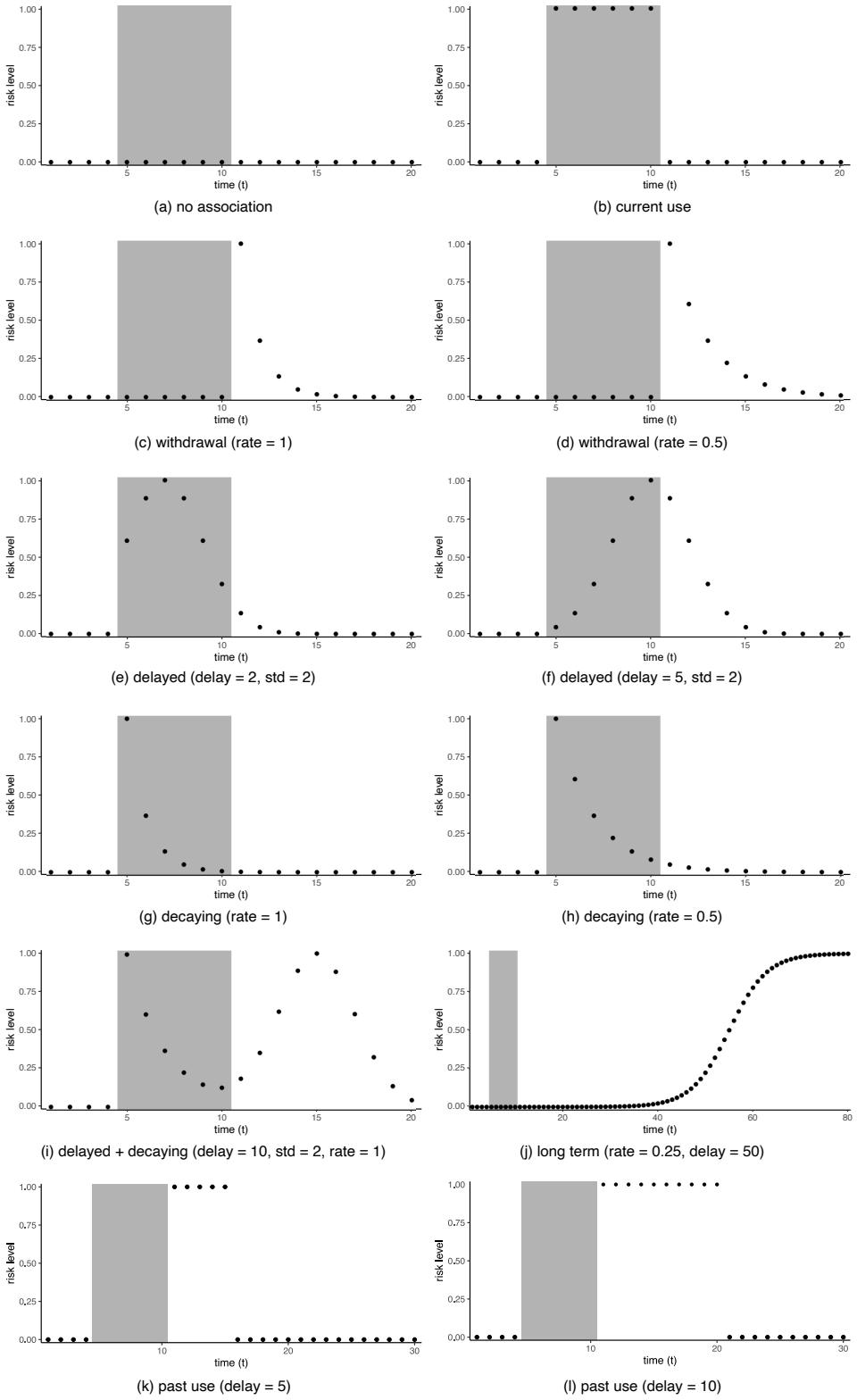

Figure 2: Examples of risk functions introduced in Section 2.3. These exact risk functions are also used in the simulation study, see Section 4. The horizontal axis represents time (t). The gray area is the period in which the patient was exposed (from $t = 5$ to $t = 10$). The y-axis shows the risk level, where 1 represents 'maximal' risk to experience the ADR, and 0 denotes 'minimal' risk.

Delayed Effects ($\mathcal{M}_{\text{delayed}}$)

The delayed effect model represents drug-ADR pairs where the risk of experiencing the ADR increases gradually during exposure, reaches a 'peak' (represented by the parameter $\mu > 0$) and dissipates afterwards. We model this using the probability density function of the normal distribution, normalized so that the risk level is 1 at μ, i.e.,

$$r_{\mathcal{M}_{\text{delayed}}}\left(\boldsymbol{X}(1:t);\mu,\sigma\right) = \begin{cases} 0 & \text{if never exposed, and} \\ \exp\left[-\frac{1}{2}\left(\frac{f_{\text{start}}(\boldsymbol{X}(1:t))-\mu}{\sigma}\right)^2\right] & \text{otherwise,} \end{cases} \tag{4}$$

for all $\boldsymbol{X} \in \mathcal{D}_T$. The parameter $\sigma > 0$ regulates how rapidly the risk level increases/decreases and the function $f_{\text{start}}(\cdot)$ returns the number of time points since the patient was exposed for the first time:

$$f_{\text{start}}\left(\boldsymbol{X}(1:t)\right) = t - \min\left\{\tau \in \{1,2,\ldots,t\} \text{ such that } X(\tau) = 1\right\}. \tag{5}$$

See Figure 2e and Figure 2f for an example where μ is 2 and 5, respectively. The parameter σ equals 2 in both cases. Various antiepilectic drugs are known to have a similar temporal relationship to the Stevens-Johnson syndrome [7]. Direct oral anticoagulants (DOACs) and gastrointestinal bleeding are also known to follow a similar pattern [19]. We consider the latter drug-ADR pair in the case study, see Section 6.

Decaying Effects ($\mathcal{M}_{\text{decaying}}$)

In case of a decaying effect, the risk level is maximal when the patient is exposed for the first time and quickly diminishes, even when still exposed. We propose

$$r_{\mathcal{M}_{\text{decaying}}}\left(\boldsymbol{X}(1:t);\rho\right) = \begin{cases} 0 & \text{if never exposed, and} \\ \exp\left(-\rho \cdot f_{\text{start}}\left(\boldsymbol{X}(1:t)\right)\right) & \text{otherwise,} \end{cases}$$

for all $\boldsymbol{X} \in \mathcal{D}_T$, where $\rho > 0$ represents the rate with which the risk level decreases. See Figure 2g and 2h for two examples where the rate ρ is either 1 or $\frac{1}{2}$. Drug-ADR pairs that are known to follow such a pattern include penicillin and anaphylaxis [34], antiepileptic drugs and adverse psychiatric effects [37, 59], and oral contraceptives and venous thrombosis [56]. We consider penicillin and anaphylaxis in the case study, see Section 6.

Delayed and Decaying Effects ($\mathcal{M}_{\text{delayed+decaying}}$)

In some cases, the exposure to a drug shows both a delayed and a decaying effect on the occurrences of the ADR, e.g., oral glucocorticoids and serious infections [15]. We can reflect this by combining both models as follows:

$$r_{\mathcal{M}_{\text{delayed+decaying}}}\left(\boldsymbol{X}(1:t);\mu,\sigma,\rho\right) = C^{-1}\left[r_{\mathcal{M}_{\text{delayed}}}\left(\boldsymbol{X}(1:t);\mu,\sigma\right) + r_{\mathcal{M}_{\text{decaying}}}\left(\boldsymbol{X}(1:t);\rho\right)\right],$$

where μ, σ and $\rho > 0$ and C is a normalizing constant. We must select the value for C such that the maximum value of the model's risk function is 1. This necessitates solving the optimization problem

$$\begin{aligned} C &= \max_{\boldsymbol{X} \in \{0,1\}^T} \left\{ r_{\mathcal{M}_{\text{delayed}}} \left(\boldsymbol{X}(1:t); \mu, \sigma \right) + r_{\mathcal{M}_{\text{decaying}}} \left(\boldsymbol{X}(1:t); \rho \right) \right\} \\ &= \max_{\boldsymbol{X} \in \{0,1\}^T} \left\{ \exp \left[-\frac{1}{2} \left(\frac{f_{\text{start}}(\boldsymbol{X}) - \mu}{\sigma} \right)^2 \right] + \exp \left[-\rho \cdot f_{\text{start}} \left(\boldsymbol{X} \right) \right] \right\}. \end{aligned}$$

To solve this, we would actually need to consider each binary vector $\boldsymbol{X} \in \{0,1\}^T$. However, we can simplify the problem using the function $f_{\text{start}}(\cdot)$. This indicates the number of time points that have elapsed since the patient was first exposed to the drug, see eq. (5). The values that the function $f_{\text{start}}(\cdot)$ can take are restricted to $1, 2, \ldots, T-1$. Consequently, we can formulate the optimization problem as

$$C = \max_{s \in \{1,2,\ldots,T-1\}} \left\{ \exp \left[-\frac{1}{2} \left(\frac{s - \mu}{\sigma} \right)^2 \right] + \exp(-\rho \cdot s) \right\},$$

which can be readily solved numerically. For an illustration of this model, see Figure 2i where $\mu = 10$, $\sigma = 2$ and $\rho = 1$.

Long-term Effects ($\mathcal{M}_{\text{long-term}}$)

The ADR can, in some cases, occur long after the patient was exposed for the first time, e.g., antipsychotics and type 2 diabetes [25]. We model these 'long-term' cases using a sigmoid function. For all $\boldsymbol{X} \in \mathcal{D}_T$,

$$r_{\mathcal{M}_{\text{long-term}}} \left(\boldsymbol{X}(1:t); \rho, \kappa \right) = \begin{cases} 0 & \text{if never exposed, and} \\ \exp(-\rho(f_{\text{start}}(\boldsymbol{X}(1:t)) - \kappa))^{-1} & \text{otherwise.} \end{cases}$$

The risk function has two parameters: $(\rho, \kappa) \in \Theta_{\mathcal{M}_{\text{long-term}}} = \mathbb{R}_+^2$. See Figure 2j for an example where $\rho = \frac{1}{4}$ and $\kappa = 50$. In contrast to the other figures, the x-axis ranges from $t = 1$ to $t = 80$ to illustrate the risk function more clearly.

Past Exposure ($\mathcal{M}_{\text{past}}$)

The risk level of the ADR in question occurring can be elevated from the start of the exposure and remain elevated for a certain period of time, even when the patient is no longer exposed. We model this as

$$r_{\mathcal{M}_{\text{past}}} \left(\boldsymbol{X}(1:t); p \right) = \begin{cases} 0 & \text{if never exposed, and} \\ 1 & \text{if } f_{\text{last}}(\boldsymbol{X}(1:t)) \leq p, \end{cases} \tag{6}$$

for all $\boldsymbol{X} \in \mathcal{D}_T$, where $p \in \{1, 2, \ldots, T-1\}$ and $f_{\text{last}}(\cdot)$ is given in eq. (3). Note that this risk function is the only one proposed here that has a discrete parameter. See Figure 2k and 2l for an example where $p = 5$ and $p = 10$, respectively. Note that in this case, the x-axis ranges from $t = 1$ to $t = 30$. An example of drug-ADR pair that approximately follows the past exposure model are antiepileptic drugs and delayed allergic hypersensitive reactions [37, 61].

2.4 Parameter Estimation

In this section, we describe how the parameters of an exposure model $\mathcal{M}$, i.e., π_0, π_1 and the parameter vector of the risk function $\boldsymbol{\theta}$, can be estimated on the basis of the data. We first consider a general exposure model. Afterwards, we derive the estimators for the no association, current use and past use exposure models as defined in the previous section since, in their case, an analytical solution exists. We denote the data for all patients as $\underset{\sim}{\boldsymbol{X}} = \{\boldsymbol{X}^k\}_{k=1}^N$ and $\underset{\sim}{\boldsymbol{Y}} = \{\boldsymbol{Y}^k\}_{k=1}^N$. The likelihood function for the exposure model $\mathcal{M}$ is given by

$$\mathcal{L}_{\mathcal{M}}\left(\pi_1, \pi_0, \boldsymbol{\theta}; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) = \prod_{k=1}^{N}\prod_{t=1}^{T}\left[\left(\pi_1 - \pi_0\right) r_{\mathcal{M}}\left(\boldsymbol{X}^k(1:t); \boldsymbol{\theta}\right) + \pi_0\right]^{Y^k(t)} \times \left[1 - \left(\pi_1 - \pi_0\right) r_{\mathcal{M}}\left(\boldsymbol{X}^k(1:t); \boldsymbol{\theta}\right) - \pi_0\right]^{1-Y^k(t)}.$$

The corresponding log-likelihood function is then

$$\begin{aligned}\ell_{\mathcal{M}}\left(\pi_1, \pi_0, \boldsymbol{\theta}; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) &= \log \mathcal{L}_{\mathcal{M}}\left(\pi_1, \pi_0, \boldsymbol{\theta}; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) \\ &= \sum_{k=1}^{N}\sum_{t=1}^{T}\Big[Y^k(t)\log\left(\left(\pi_1 - \pi_0\right) r_{\mathcal{M}}\left(\boldsymbol{X}^k(1:t); \boldsymbol{\theta}\right) + \pi_0\right) + \\ &\qquad \left(1 - Y^k(t)\right)\log\left(1 - \left(\pi_1 - \pi_0\right) r_{\mathcal{M}}\left(\boldsymbol{X}^k(1:t); \boldsymbol{\theta}\right) - \pi_0\right)\Big].\end{aligned} \tag{7}$$

Recall that the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\psi})$ can consist of both continuous ($\boldsymbol{\xi}$) and discrete parameters ($\boldsymbol{\psi}$). The maximum likelihood estimator (MLE) can, therefore, be written as

$$\begin{aligned}\left(\widehat{\pi}_1, \widehat{\pi}_0, \widehat{\boldsymbol{\theta}}\right) &= \underset{\pi_1, \pi_0 \in [0,1], \boldsymbol{\xi} \in \Xi_{\mathcal{M}}, \boldsymbol{\psi} \in \Psi_{\mathcal{M}}}{\arg\max}\ \ell_{\mathcal{M}}\left(\pi_1, \pi_0, (\boldsymbol{\xi}, \boldsymbol{\psi}); \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) \\ &= \underset{\boldsymbol{\psi} \in \Psi_{\mathcal{M}}}{\arg\max}\left\{\underset{\pi_1, \pi_0 \in [0,1], \boldsymbol{\xi} \in \Xi_{\mathcal{M}}}{\arg\max}\ \ell_{\mathcal{M}}\left(\pi_1, \pi_0, (\boldsymbol{\xi}, \boldsymbol{\psi}); \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right)\right\}.\end{aligned}$$

In other words, the original optimization problem can be subdivided into $|\Psi_{\mathcal{M}}|$ subproblems (where $|\cdot|$ denotes the cardinality of the set), one for each value $\psi \in \Psi_{\mathcal{M}}$. Analytical solutions for these subproblems might not exist, but they can be solved numerically. We employ Nelder-Mead's algorithm since it allows for discontinuous risk functions as well [33, 5].

No Association ($\mathcal{M}_0$)

Recall that the risk function for the null model is $r_{\mathcal{M}_0}(\boldsymbol{X}) = 0$ for all $\boldsymbol{X} \in \mathcal{D}_T$. Let $Y^+ = \sum_{k=1}^{N} \sum_{t=1}^{T} Y^k(t)$ be the total number of occurrences of the ADR in the data set. The log-likelihood function reduces to

$$\begin{aligned}\ell_{\mathcal{M}_0}\left(\pi_0; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) &= \sum_{k=1}^{N}\sum_{t=1}^{T} Y^k(t)\log\left(\pi_0\right) + \left(1 - Y^k(t)\right)\log(1-\pi_0) \\ &= Y^+ \log\left(\pi_0\right) + \left(NT - Y^+\right)\log\left(1-\pi_0\right).\end{aligned}$$

Maximizing this function with respect to π_0 gives the MLE $\widehat{\pi}_0 = \frac{Y^+}{NT}$.

Current use model ($\mathcal{M}_{\text{current use}}$)

We can determine the MLE for this model in a similar fashion. Let us first define the following 2×2 contingency table represented by the random variables A, B, C and D, which are given by

$$A = \sum_{k=1}^{N}\sum_{t=1}^{T} X^k(t)Y^k(t), \quad B = \sum_{k=1}^{N}\sum_{t=1}^{T} X^k(t)(1 - Y^k(t)),$$
$$C = \sum_{k=1}^{N}\sum_{t=1}^{T}(1 - X^k(t))Y^k(t) \quad \text{and} \quad D = \sum_{k=1}^{N}\sum_{t=1}^{T}(1 - X^k(t))(1 - Y^k(t)).$$

The count A represent the number of time points the patients were exposed to the drug and experienced the ADR, B is the number of times the patient was exposed, but did not experience the ADR, etc. Note that the sum of these counts are the total number of observed time points, i.e., $A+B+C+D = NT$. Using these definitions, we can express the exposure model's log-likelihood as

$$\ell_{\mathcal{M}_{\text{current use}}}\left(\pi_1, \pi_0; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) = A\log\left(\pi_1\right) + B\log\left(1-\pi_1\right) + C\log\left(\pi_0\right) + D\log\left(1-\pi_0\right).$$

If $A + B > 0$ and $C + D > 0$, the MLE can be calculated as

$$(\widehat{\pi}_1, \widehat{\pi}_0) = \left(\frac{A}{A+B}, \frac{C}{C+D}\right).$$

Past Exposure ($\mathcal{M}_{\text{past}}$)

In order to derive the MLE for the past use model, we first define the counts $A(p)$, $B(p)$, $C(p)$ and $D(p)$ for $p = 1, 2, \ldots, T-1$ as follows:

$$
\begin{aligned}
A(p) &= \sum_{k=1}^{N}\sum_{t=1}^{T} Y^k(t)\mathbb{1}\left\{\exists\tau \in \{\max\{1, t-p\}, \ldots, t\} \text{ such that } X^k(t) = 1\right\}, \\
B(p) &= \sum_{k=1}^{N}\sum_{t=1}^{T} \left(1 - Y^k(t)\right)\mathbb{1}\left\{\exists\tau \in \{\max\{1, t-p\}, \ldots, t\} \text{ such that } X^k(t) = 1\right\}, \\
C(p) &= \sum_{k=1}^{N}\sum_{t=1}^{T} Y^k(t)\mathbb{1}\left\{\forall\tau \in \{\max\{1, t-p\}, \ldots, t\} : X^k(t) = 0\right\} \text{ and} \\
D(p) &= \sum_{k=1}^{N}\sum_{t=1}^{T} \left(1 - Y^k(t)\right)\mathbb{1}\left\{\forall\tau \in \{\max\{1, t-p\}, \ldots, t\} : X^k(t) = 0\right\},
\end{aligned}
$$

where $A(p)$ denotes the number of occurrences of the ADR when the patient was exposed during the last p time points, $B(p)$ denotes the number of time points when the patient did not experience the ADR, but was exposed during the last p time points etc. We can express the log-likelihood function of this exposure model for a fixed p as

$$
\ell_{\mathcal{M}_{\text{past}}}\left(\pi_1, \pi_0, p; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) = A(p)\log\left(\pi_1\right) + B(p)\log\left(1-\pi_1\right) + C(p)\log\left(\pi_0\right) + D(p)\log\left(1-\pi_0\right)
$$

which gives the following MLE of (π_1, π_0), if $A(p) + B(p) > 0$ and $C(p) + D(p) > 0$:

$$
(\widehat{\pi}_1, \widehat{\pi}_0) = \left(\frac{A(p)}{A(p)+B(p)}, \frac{C(p)}{C(p)+D(p)}\right).
$$

Using the result in eq. (7), the estimator for the past use model can be expressed as

$$
\begin{aligned}
(\widehat{\pi}_1, \widehat{\pi}_0, \widehat{p}) &= \underset{p\in\{1,2\ldots,T-1\}}{\arg\max} \left\{ \underset{\pi_1,\pi_0\in[0,1]}{\arg\max}\, \ell_{\mathcal{M}_{\text{past}}}\left(\pi_1, \pi_0, p; \underset{\sim}{\boldsymbol{X}}, \underset{\sim}{\boldsymbol{Y}}\right) \right\} \\
&= \underset{p\in\{1,2\ldots,T-1\}}{\arg\max} \left\{ A(p)\log\left(\frac{A(p)}{A(p)+B(p)}\right) + B(p)\log\left(\frac{B(p)}{A(p)+B(p)}\right) + \right. \\
&\qquad \left. C(p)\log\left(\frac{C(p)}{C(p)+D(p)}\right) + D(p)\log\left(\frac{D(p)}{C(p)+D(p)}\right) \right\}.
\end{aligned}
$$

2.5 Model Selection

There is a variety of model selection approaches available in the literature that can aid in selecting the 'best' exposure model after fitting it to the data [14]. Here we opt for the

Bayesian Information Criterion (BIC) due to its connection to the posterior probabilities of the models [43]. The BIC for an exposure model $\mathcal{M}$ is given by

$$\text{BIC} = (q+2)\log(N) - 2\ell_{\mathcal{M}}\left(\pi_1^*, \pi_0^*, \boldsymbol{\theta}^*; \underset{\sim}{\underset{\sim}{\boldsymbol{x}}}, \underset{\sim}{\boldsymbol{y}}\right),$$

where q is the number of parameters of the risk function, i.e., the dimensionality of the parameter space $\Theta_{\mathcal{M}}$, $\underset{\sim}{\underset{\sim}{\boldsymbol{x}}} = \{\boldsymbol{x}^k\}_{k=1}^N$ and $\underset{\sim}{\boldsymbol{y}} = \{\boldsymbol{y}^k\}_{k=1}^N$ are the observed data, and π_1^*, π_0^* and $\boldsymbol{\theta}^*$ are the values for which the log-likelihood is maximal. Note that with the BIC, one tries to strike a balance between the fit (log-likelihood) and the model's complexity expressed by the total number of parameters $(q+2)$.

Suppose we consider the exposure models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_V$, where $V = 8$ in our case, and let BIC_v be the BIC-score for the v-th model. The model with the lowest BIC-score is preferred. Schwarz[43] shows that the posterior probability of the v-th model can be approximated by

$$\mathbb{P}\left(\mathcal{M}_v \mid \underset{\sim}{\underset{\sim}{\boldsymbol{x}}}, \underset{\sim}{\boldsymbol{y}}\right) \approx \frac{\exp\left(-\frac{1}{2}\text{BIC}_v\right)}{\sum_{w=1}^{V}\exp\left(-\frac{1}{2}\text{BIC}_w\right)}.$$

This result is especially useful in our case for two reasons. First, one might be interested in determining whether there is an association between the drug and ADR in question and not in identifying which exposure model fits the data 'best' in itself. To this end, one can use the posterior probability of the no association exposure model. For example, if the posterior probability $\mathbb{P}(\mathcal{M}_0 \mid \underset{\sim}{\underset{\sim}{\boldsymbol{x}}}, \underset{\sim}{\boldsymbol{y}}) \geq \frac{1}{2}$, there is no association and the drug-ADR pair is not reported as a signal.

The second reason why the use of the posterior probabilities is convenient, is its ability to facilitate a comparison of various drug-ADR pairs and establish a ranking from 'interesting' to less 'interesting'. Let $\underset{\sim}{\boldsymbol{x_i}}$ and $\underset{\sim}{\boldsymbol{y_j}}$ be the observed data for the i-th drug and the j-th ADR, respectively, and let $\mathbb{P}(\mathcal{M}_0^{ij} \mid \underset{\sim}{\boldsymbol{x_i}}, \underset{\sim}{\boldsymbol{y_j}})$ be the posterior probability of the null model for the drug-ADR pair (i,j). We can create a ranking of pairs based on these scores, where the lower the posterior probability is, the stronger the signal for the pair is deemed to be. An advantage of this approach is that one can employ Bayesian false discovery rate control procedures, see, for example, the work by Storey[46], to determine which signals to present to the committee of medical experts.

3 Simulating Electronic Healthcare Data

In this section, we describe the simulator for EHC data used for the simulation study. We start with how we model drug exposures over time, after which we show how to generate ADR occurrences given the drug exposure history and an exposure model. We finish with the pseudo-algorithm.

Drug Exposures

We model the exposure of a patient to the drug of interest over time as a Markov chain. Let $\boldsymbol{X} = (X(1), X(2), \ldots, X(T))$ be the binary time series. We then specify the Markov chain by

$$\mathbb{P}(X(1) = 1) = \nu_0 \quad \text{and} \quad \mathbb{P}(X(t) = 1 \mid X(t-1) = x) = \begin{cases} \nu_0 & \text{if } x = 0 \\ \nu_1 & \text{if } x = 1 \end{cases} \tag{8}$$

for $t = 2, 3, \ldots, T$. Rather than thinking in terms of the probabilities ν_0 and ν_1, we find it more intuitive to consider 1) the probability of the patient to be exposed to the drug at least once, and 2) the average duration of the exposure once exposed. Let E be a binary random variable representing whether the patient was exposed at least once ($E = 1$) or not ($E = 0$), and let $D \in \mathbb{N}$ be a random variable denoting the duration of the exposure once exposed. We find that the probability of being exposed can be expressed as

$$\mu_E = \mathbb{P}(E = 1) = 1 - \mathbb{P}(E = 0) = 1 - \mathbb{P}(X(1) = 0, X(2) = 0, \ldots, X(T) = 0) = 1 - (1 - \nu_0)^T .$$

So, rather than choosing ν_0 directly, we select the probability of a patient to be exposed (μ_E) and set $\nu_0 = 1 - (1 - \mu_E)^{\frac{1}{T}}$.

Once a patient is exposed, the probability that he/she is exposed at the following time point is ν_1. The duration D of the exposure, therefore, follows a geometric distribution with probability density function

$$\mathbb{P}(D = d) = (\nu_1)^d (1 - \nu_1)$$

with mean

$$\delta = \mathbb{E}(D) = (1 - \nu_1)^{-1}.$$

Once one chooses the average duration of the exposure, δ, one can set $\nu_1 = (\delta - 1)/\delta$. In our simulation set-up, we choose an average duration of $\delta = 5$ time points, i.e., $\nu_1 = .8$.

Adverse Drug Reactions

Let $\boldsymbol{X} = \boldsymbol{x} = (x(1), x(2), \ldots, x(T))$ be a simulated drug history. Given an exposure model $\mathcal{M}$ with risk function $r_{\mathcal{M}}(\cdot; \boldsymbol{\theta})$ and probabilities π_0 and π_1, the random binary variable $Y(t)$ follows, conditional on the drug history, a Bernoulli distribution, i.e.,

$$Y(t) \mid \boldsymbol{X}(1:t) = \boldsymbol{x}(1:t) \sim \text{Bernoulli}\left((\pi_1 - \pi_0) r_{\mathcal{M}}(\boldsymbol{x}(1:t); \boldsymbol{\theta}) + \pi_0\right). \tag{9}$$

Pseudo-Algorithm

Combining these steps, we propose the following procedure for generating EHC data for a single drug-ADR pair:

1. Select the number of patients (N), an exposure model $\mathcal{M} = \langle \pi_1, \pi_0, T, r_{\mathcal{M}}(\cdot; \boldsymbol{\theta}), \Theta \rangle$, the probability of a patient to be exposed (μ_E), and the average duration of the exposure once exposed (δ);

2. Determine the probabilities $\nu_0 = 1 - (1 - \mu_E)^{\frac{1}{T}}$ and $\nu_1 = (\delta - 1)/\delta$ governing the Markov chain in eq. (8);

3. For all patients $k = 1, 2, \ldots, N$ perform the following two steps:

   (a) Sample a drug history for patient k according to eq. (8). We denote the resulting drug history by $\boldsymbol{x}^k$, and

   (b) Generate the ADR history $\boldsymbol{y}^k = (y^k(1), y^k(2), \ldots, y^k(T))$ for patient k by sampling from the Bernoulli distribution in eq. (9) given the drug history $\boldsymbol{x}^k$ from the previous step.

An implementation of this algorithm is publicly available as an `R` package at `https://github.com/bips-hb/expard`.

4 Simulation Set-Up

In our simulation, we simulate one drug-ADR pair at the time following the procedure as described in the previous section. The drug-ADR pair can follow one out of twelve exposure models. We list them here. For a visual representation, see Figure 2;

1. no association;

2. current use;

3. a withdrawal effect with a rate of $\rho = 1$;

4. a withdrawal effect with a rate of $\rho = \frac{1}{2}$;

5. a delayed effect with its 'peak' at time point $\mu = 2$, and $\sigma = 2$;

6. a delayed effect with its 'peak' at time point $\mu = 5$, and $\sigma = 2$;

7. a decaying effect with a rate of $\rho = 1$;

8. a decaying effect with a rate of $\rho = \frac{1}{2}$;

9. a combination of a delayed and decaying effect where $(\mu, \sigma, \rho) = (10, 2, 1)$;

10. a long-term effect with rate $\rho = \frac{1}{4}$ and $\kappa = 50$;

11. a past use model with $p = 5$, and

12. a past use model with $p = 10$.

The number of patients and time points are fixed at $N = 1,000$ and $T = 100$, respectively. The probability for a patient to be exposed to the drug (μ_E) is varied from .01, .1 to .5. The average duration of an exposure once exposed (δ) is 5 time points. The probability π_1 is .01, .1, .2 or .3; the probability π_0 is either 10^{-4} or 10^{-3}. See Table 1 for an overview of all the parameter settings. We, thus, consider a total of 288 parameter settings. We repeat the simulation for each setting 20 times.

Table 1: Parameter settings used in the EHC simulation study

Description	**Notation**	**Values**
Number of patients	N	1000
Number of time points	T	100
Probability being exposed	μ_E	.01, .1 or .5
Average duration exposure	δ	5
Probability ADR minimal risk	π_0	10^{-4} or 10^{-3}
Probability ADR maximal risk	π_1	.01, .1, .2 or .3

5 Performance Assessment

We assess the method's effectiveness by considering two aspects: 1) its ability to accurately identify the correct exposure model (model selection), and 2) its capacity to ascertain the presence or absence of an association between the drug and the ADR (signal detection). Although these two tasks are related, it is important to note that it might be of interest to understand the nature of the relationship between the drug and ADR or to simply confirm the existence of a relationship. In the following two sections, we detail how we measure the method's performance with regard to these aspects.

5.1 Model Selection

We consider two approaches for model selection. We either select 1) the model with the lowest BIC (or, conversely, the highest posterior probability), see Section 2.5, or 2) the model with the highest likelihood. The former approach considers the complexity of the

models in terms of the number of parameters, while the latter solely focuses on the model's fit.

For every combination of μ_E, π_0, and π_1 (see Section 4), we evaluate the performance by examining the *confusion matrix*. Examples of such a confusion matrix can be found in Figures 3, 4 and 5. The horizontal axis represents the twelve true exposure models (see Section 4 for an overview), while the vertical axis indicates the model selected based on the BIC. The values within each cell indicate the number of times the respective selected model was chosen when the data was simulated based on the corresponding true model. For readability, cells with the value 0 are left empty. The difference between the number of true models (twelve) and selectable models (eight) is due to the fact that true models serve as the foundation for the simulation with predetermined parameters, whereas the parameters of the chosen model are estimated based on the data (see Section 2.4). Note that each column in the confusion matrix always adds up to 20, corresponding to the number of repetitions. To clarify, Figure 3 presents the ideal confusion matrix, where each true model is correctly identified. For example, withdrawal models with rates 1 and $\frac{1}{2}$ are both identified as the withdrawal model.

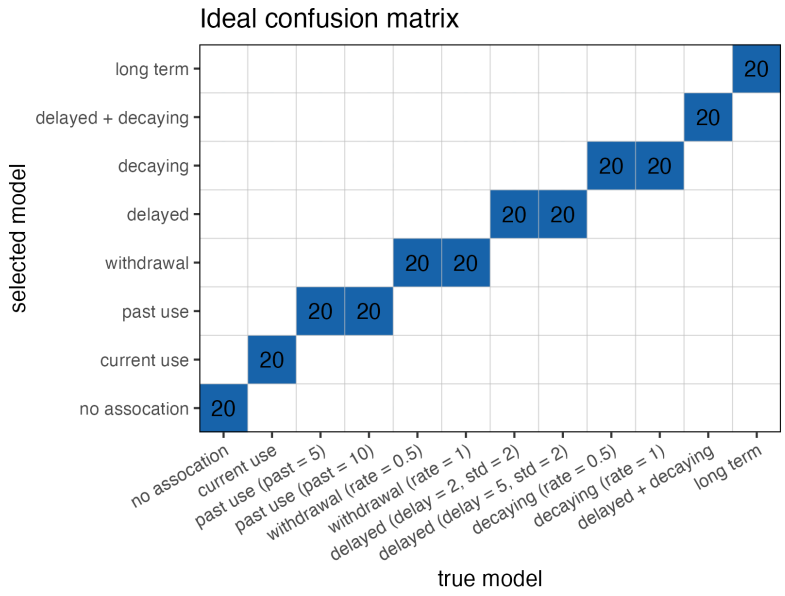

Figure 3: The confusion matrix in the ideal scenario where the method classifies each case perfectly. The twelve true models utilized in the simulation study (see Section 4) are represented along the horizontal axis. On the vertical axis, the exposure model that was selected is shown. Note that, for instance, choosing 'withdrawal' is correct for both true withdrawal models where the rate is either 1 or $\frac{1}{2}$. The number 20 denotes the total number of repetitions.

5.2 Signal Detection

Our focus here lies in evaluating the the method's performance in detecting the presence or absence of an association between the drug and ADR of interest. Recall that we simulate twelve exposure models for every combination of μ_E, π_0, and π_1, as outlined in Section 4. We repeat this process 20 times, resulting in a total of $12 \cdot 20 = 240$ runs. A drug-ADR pair is not associated when the true model is the no association model $\mathcal{M}_0$, and associated if the true model is one of the remaining eleven. To frame this as a binary classification problem, we define two sets: let $\boldsymbol{I}^* = \{I_1^*, I_2^*, \ldots, I_{240}^*\}$ represent the truth, where $I_l^* = 0$ if, for the l-th run, the true model is the null model, and 1 otherwise. Note that only 20 values in $\boldsymbol{I}^*$ are 0. Let $\boldsymbol{I} = \{I_1, I_2, \ldots, I_{240}\}$ represent the method's decisions, where I_l is 0 if the method indicates no association, and 1 otherwise. We can then define the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as

$$\text{TP} = \sum_{l=1}^{240} I_l^* I_l, \quad \text{TN} = \sum_{l=1}^{240} \left(1 - I_l^*\right)\left(1 - I_l\right),$$
$$\text{FP} = \sum_{l=1}^{240} \left(1 - I_l^*\right) I_l \quad \text{and} \quad \text{FN} = \sum_{l=1}^{240} I_l^* \left(1 - I_l\right).$$

In situations involving unbalanced data, as is the case here, it is recommended to use precision and recall as performance measures [40]. These metrics are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and} \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

The F_1 score is their harmonic mean, i.e.,

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{11}$$

To highlight the impact of using either the posterior probability (linked to the BIC) or the likelihood directly, we decide on the presence of an association between the drug and ADR based on whether: 1) the posterior probability of the null model exceeds .5, or 2) the model with the highest likelihood is not the null model.

6 Case Study

To demonstrate our exposure model framework, we implement the suggested approach using data from the German Pharmacoepidemiological Research Database (GePaRD; [23]). The data is from two statutory health insurance (SHI) providers in Germany: hkk Krankenkasse and AOK Bremen/Bremerhaven. Our analysis focuses solely on in-patient data and individuals who were insured 1) during the years 2004 until 2017, and 2) for a consecutive period

of time, allowing for a maximum gap of 14 days. The total number of patients considered exceeds 1.2 million. The temporal resolution is set to quarter years.

We consider four drug-ADR pairs for which the temporal relationship is known in the literature:

1. Penicillin (ATC: J01C) and anaphylaxis (ICD: T88.6): In the case of a penicillin allergy, the reaction typically occurs almost immediately, with the associated risk diminishing rapidly over time. This aligns closely with the decaying exposure model [34]. However, it is important to note that detecting this pattern requires high time resolution (days rather than quarters). Consequently, the current use model is anticipated to offer a more fitting description of the relationship;

2. Direct oral anticoagulants (DOACs; ATC: B01AF) and gastrointestinal (GI) bleeding[†]: The probability of bleeding increases with prolonged exposure, reaches a 'peak' and decreases afterwards. This pattern aligns most closely with the delayed exposure model [19];

3. Antipsychotics (ATC: N05A) and type 2 diabetes (ICD: E11): Prolonged use of antipsychotics has been associated with weight gain and the onset of type 2 diabetes. The long-term exposure model seems to be most appropriate [25], and

4. Antibiotics (ATC: J01) and GI bleeding: This combination is our negative control, as there is no evidence that the use of antibiotics elevates the risk of GI bleeding. We incorporate this negative control to evaluate the effectiveness of our method in identifying the null model.

7 Results

We present the results from both the simulation and the case study.

7.1 Simulation Study

Initially, we investigate the method's capability to select the true exposure model. Subsequently, we evaluate its effectiveness in accurately determining the existence of an association between the drug and the ADR.

[†]The ICD-codes related to gastrointestinal (GI) bleeding are: I98.3, K22.6, K22.8, K22.80, K22.81, K22.88, K25.0, K25.2, K25.4, K25.6, K26.0, K26.2, K26.4, K26.6, K27.0, K27.2, K27.4, K27.6, K28.0, K28.2,K 28.4, K28.6, K29.0, K31.8, K55.2, K55.3, K55.8, K57.0, K57.1, K57.2, K57.3, K57.4, K57.5, K57.8, K57.9, K62.5, K66.1, K92.0, K92.1, K92.2.

7.1.1 Model Selection

In this section, we show the simulation results for a select number of parameter settings. You can interactively explore the results for the other parameter settings at `https://exposuremodels.bips.eu`. The trends shown here in this section are applicable to the broader range of parameter settings as well.

Figures 4 and 5 display a set of confusion matrices (see Section 5.1), where the probability of being exposed to the drug at least once (μ_E) is, respectively, relatively low, i.e., around 1% of the patients, and high, i.e., around 50%. Both figures are organized as follows: rows represent different values of π_0, signifying the probability of experiencing the ADR when the patient is at minimal risk (a risk level of 0). The top and bottom rows correspond to $\pi_0 = 10^{-4}$ and $\pi_0 = 10^{-3}$, respectively. Columns represent different values of π_1, denoting the probability of the ADR occurring when the patient is at maximal risk (a risk level of 1). The left and right columns correspond to $\pi_1 = .01$ and $\pi_1 = .3$, respectively.

Figure 5 illustrates the simulation results when the probability of exposure is high ($\mu_E = .5$). The figure follows the same structure as Figure 4, with the rows presenting the results for π_0 values of 10^{-4} and 10^{-3}, respectively, and with the columns containing the results for π_1 values of .01 and .3, in that order.

The lowest performance observed shows the confusion matrix in the lower left corner of the figure. Two factors contribute to the poor performance: 1) the probability of experiencing the ADR at maximal risk, denoted as π_1, is relatively low ($\pi_1 = .01$), and 2) the difference between the probabilities of experiencing the ADR at maximal and minimal risk, represented as $|\pi_1 - \pi_0|$, is smaller compared to the top row of the figure. A reduced disparity between π_1 and π_0 makes it more challenging to detect an association between the drug and the specific ADR. While the performance in the top left corner is slightly better, reliably identifying a signal under these conditions remains impossible. The suboptimal results in the left column of the figure are not surprising; only a limited number of patients are exposed, and even when individuals are at risk, occurrences of the ADR are infrequent.

The outcomes in the right column of Figure 4 show significant improvement, mainly because the difference $|\pi_1 - \pi_0|$ is considerably larger than in the left column. The probability of the ADR occurring is significantly higher when the patient is at risk, making it easier to accurately identify the exposure model.

The exposure models delayed or delayed + decaying, see Section 2.3, are frequently misclassified as the past use model. This misclassification arises due to the fact that these models share important characteristics; namely, the delayed, delayed + decaying, and past use models all indicate that the patient is at an increased risk of experiencing the ADR during or after being exposed to the drug, see Figure 2. Given our use of the BIC for model selection, the past use model is favored due to its fewer parameters – three as opposed to

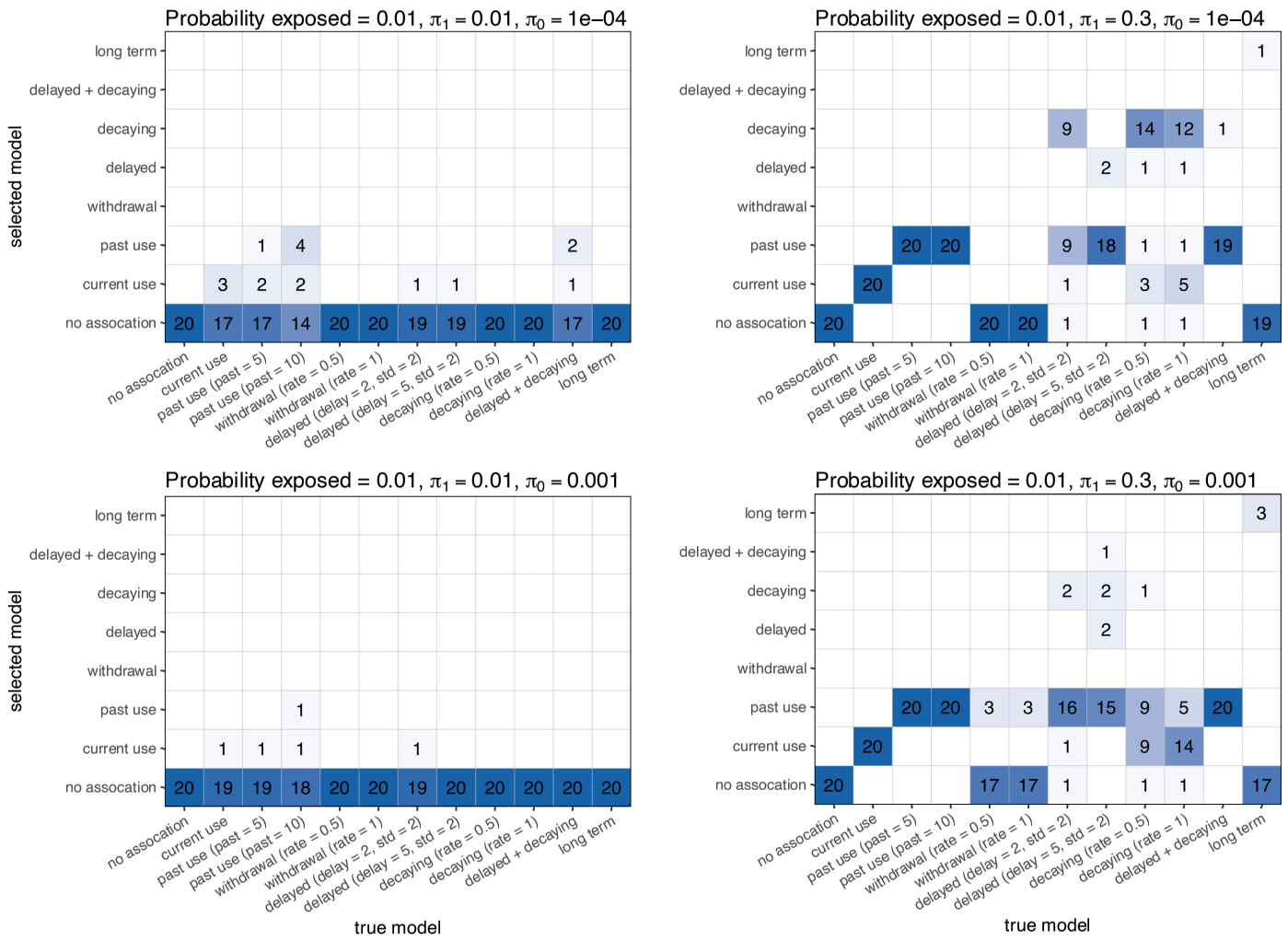

Figure 4: Several confusion matrices showing the simulation results when the number of exposed patients is low, approximately 1% of the patients. The x-axes represent the twelve true models, while the y-axes indicate the models selected based on the BIC. The top and bottom rows illustrate the results when π_0 is 10^{-4} and 10^{-3}, respectively. The left and right columns present the outcomes for π_1 values of .01 and .3, respectively.

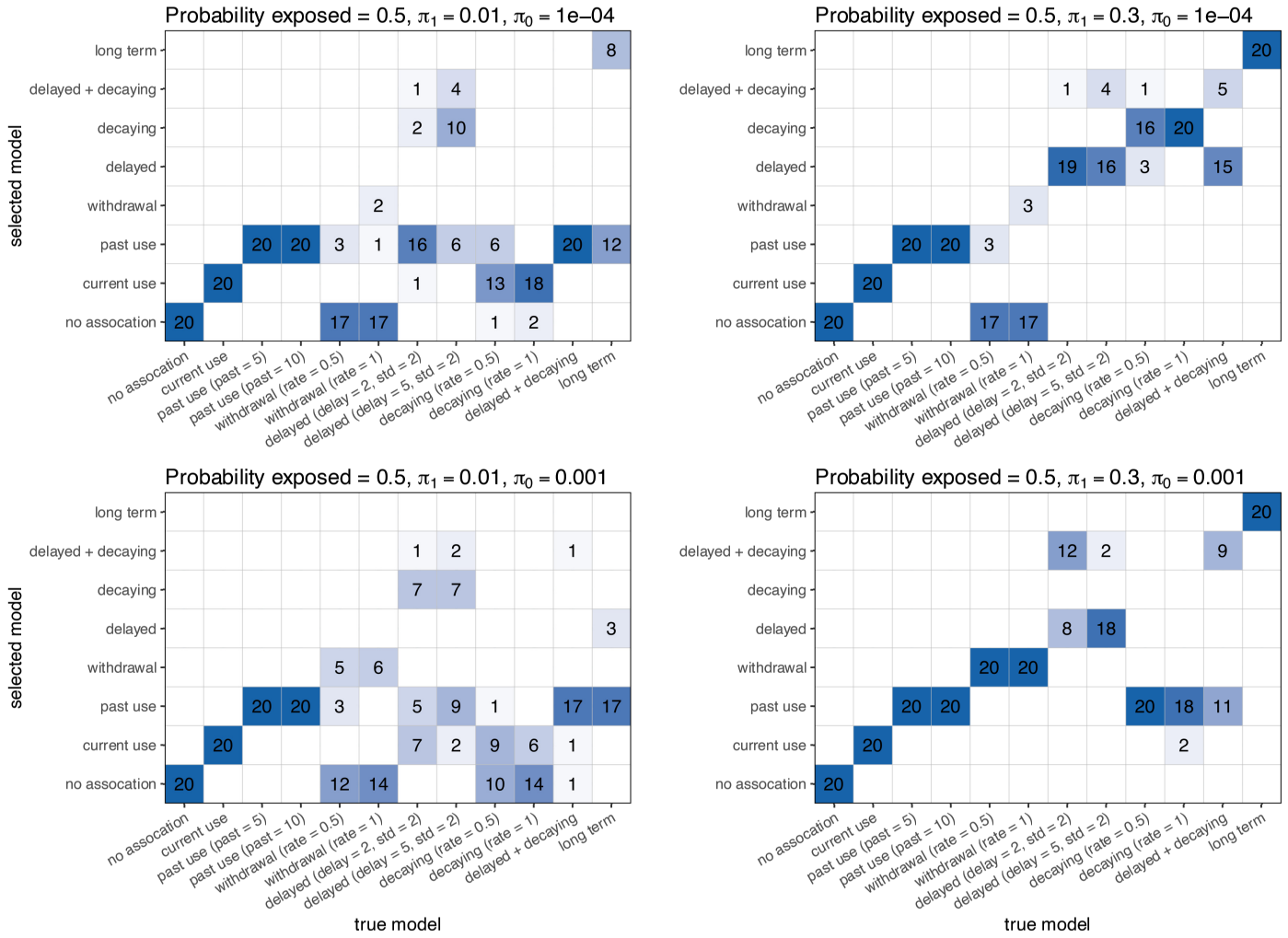

Figure 5: Several confusion matrices showing the simulation results when the number of exposed patients is high, approximately 50% of the patients. The x- and y-axes denote the true and selected models, respectively. The top and bottom rows illustrate the results with π_0 set at 10^{-4} and 10^{-3}, while the left and right columns present the outcomes for π_1 values of .01 and .3, in that order.

four and five parameters for the delayed and delayed + decaying models, see Section 2.3. The available data do not provide sufficient support for differentiating between these three models. However, this scenario changes in Figure 5 where the number of exposed patients is significantly higher.

For most exposure models, detection improves as the difference $|\pi_1 - \pi_0|$ increases. However, this trend does not apply to the withdrawal model, which, surprisingly, becomes easier to detect when the difference decreases slightly. Although this effect is less pronounced in this figure, it becomes more evident in Figure 5.

Detecting long-term models proves to be very challenging under the settings considered.

Let us consider the results in Figure 5 when the probability of exposure is high ($\mu_E = .5$). The performance significantly improves compared to the scenario depicted in Figure 4. This improvement is expected since approximately half of the patients are now exposed, a substantial increase from the previous 1%. Similarly, we observe that the performance is influenced by the difference between the probability of the ADR occurring when the patient is at maximal or minimal risk, i.e., $|\pi_1 - \pi_0|$. Specifically, the top row, where this difference is larger, exhibits better performance than the bottom row.

Furthermore, the delayed and decaying models and their combination are still frequently misclassified as the past use model for the same reason mentioned earlier. The optimal performance occurs in the upper right corner, indicating that under these conditions, the data supports choosing these models over the past use model. The withdrawal model is easier to detect when the difference $|\pi_1 - \pi_0|$ is smaller, contrary to the other models. However, the long-term model remains challenging to detect when π_1 is .01 (left column) but becomes detectable when π_1 is .3 (right column).

7.1.2 Signal Detection

The results are presented in Table 2, where the first three columns display the parameter settings, the next three columns showcase the results when the posterior probability is utilized, and the last three columns illustrate the outcomes when the decision is based on the likelihood of the model.

Similar patterns emerge here as those observed in the previous section on model selection. Performance tends to be poor when the number of patients exposed at least once is low, improving as this number increases. As the probability of experiencing the ADR at maximal risk (π_1) increases, performance also improves, aligning with expectations. Comparing cases where μ_E and π_1 remain constant show better performance when $\pi_0 = 10^{-4}$ is low, since the difference between π_1 and π_0 is larger.

Utilizing the posterior probability results in overly conservative decisions, as evidenced by a precision of 1 for all parameter settings, indicating a reluctance to produce a signal.

The use of maximum likelihood improves the situation, although precision may decrease in some instances. Notably, when there are more patients exposed or the frequency of the ADR increases, the performance of the method based on the posterior probability surpasses that of the maximum likelihood-based approach.

Table 2: The performance when either the posterior probability or the maximum likelihood is used for determining whether there is an association between the drug and ADR in question. The parameters used in the simulation are the probability to be exposed (μ_E) and the probabilities for the ADR to occur when the patient is at minimal (π_0) or maximal risk (π_1). See for the definitions of the precision, recall and F_1 score eq. (10) and (11).

Parameter settings			**Posterior probability**			**Max. likelihood**		
μ_E	π_0	π_1	**Precision**	**Recall**	F_1	**Precision**	**Recall**	F_1
.01	10^{-4}	.01	1	.08	.14	1	.22	.36
.01	10^{-4}	.10	1	.50	.67	1	.58	.74
.01	10^{-4}	.20	1	.70	.82	1	.71	.83
.01	10^{-4}	.30	1	.72	.84	1	.73	.85
.01	10^{-3}	.01	1	.02	.04	.95	.27	.42
.01	10^{-3}	.10	1	.45	.62	.98	.62	.76
.01	10^{-3}	.20	1	.69	.81	.98	.75	.85
.01	10^{-3}	.30	1	.75	.86	.98	.77	.86
.10	10^{-4}	.01	1	.53	.69	1	.64	.78
.10	10^{-4}	.10	1	.75	.86	1	.75	.86
.10	10^{-4}	.20	1	.77	.87	1	.77	.87
.10	10^{-4}	.30	1	.78	.88	1	.78	.88
.10	10^{-3}	.01	1	.27	.43	.95	.77	.85
.10	10^{-3}	.10	1	.87	.93	.95	.87	.91
.10	10^{-3}	.20	1	.87	.93	.96	.87	.91
.10	10^{-3}	.30	1	.88	.93	.96	.88	.91
.50	10^{-4}	.01	1	.83	.91	.98	.85	.91
.50	10^{-4}	.10	1	.85	.92	.98	.85	.91
.50	10^{-4}	.20	1	.85	.92	.98	.85	.91
.50	10^{-4}	.30	1	.85	.92	.98	.85	.91
.50	10^{-3}	.01	1	.77	.87	.92	1	.96
.50	10^{-3}	.10	1	1	1	.92	1	.96
.50	10^{-3}	.20	1	1	1	.92	1	.96
.50	10^{-3}	.30	1	1	1	.92	1	.96

7.2 Case Study

This section presents the outcomes of the case study (see Section 6), where the proposed method was applied to the four drug-ADR pairs introduced above. Three of these drug-ADR pairs are known to be associated, while the combination of antibiotics and GI bleeding serves as our negative control. The data set contains information on over 1.2 million insurants from 2004 until 2017. Given that the data are in quarter-years and we consider a time period of 14 years, there are a total of $T = 56$ time points.

We treat drug dispensation as synonymous with drug exposure. This is a strong assumption for several reasons, e.g., it assumes perfect adherence and overlooks potential changes in the treatment plan.

Table 3 displays four 2×2 contingency tables, one for each drug-ADR pair. In each table, the entries represent the number of individuals who were dispensed the drug at least once and/or experienced the ADR at least once. For example, the observed number of patients who were dispensed penicillin at any time and experienced anaphylactic shock during their coverage period is 25. Similarly, the count of patients who experienced anaphylactic shock throughout their observed period but were not dispensed penicillin is 171, and so forth.

Table 3: The 2×2 contingency tables for the four drug-ADR pairs considered in the case study. Each table shows the number of patients that were dispensed the drug and/or experienced the ADR.

	ADR	not ADR	*total*
drug	25	74,068	74,093
not drug	171	1,179,093	1,179,264
total	196	1,253,161	1,253,357

(a) Penicillin and anaphylaxis

	ADR	not ADR	*total*
drug	852	16,111	16,963
not drug	10,252	1,226,142	1,236,394
total	11,104	1,242,253	1,253,357

(b) DOACs and GI bleeding

	ADR	not ADR	*total*
drug	2,079	59,515	61,594
not drug	6,634	1,185,129	1,191,763
total	8,713	1,244,644	1,253,357

(c) Antipsychotics and type 2 diabetes

	ADR	not ADR	*total*
drug	9,082	716,990	726,072
not drug	2,022	525,263	527,285
total	11,104	1,242,253	1,253,357

(d) Antibiotics and GI bleeding

Figure 6 shows the BIC scores for all eight exposure models for the four drug-ADR pairs. The models are arranged from the best (based on the BIC) on the left to the worst on the right. In all figures, except for the lower right plot, the BIC value for the no association

model is notably higher than the other BIC scores, surpassing the limits of the y-axis. We have included the rounded value of the BIC in white on the corresponding bars. We discuss each drug-ADR pair individually.

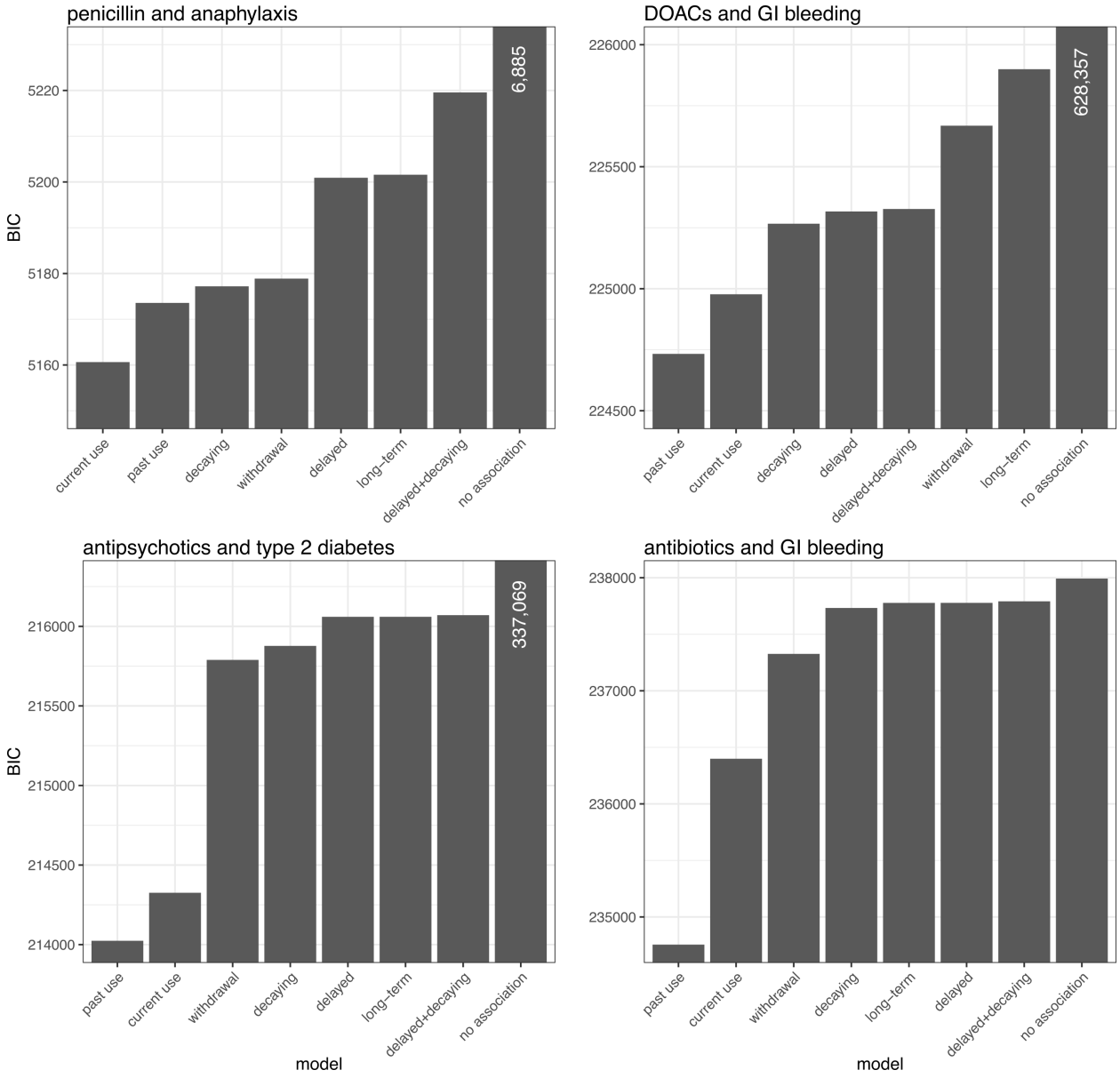

Figure 6: The BIC values for the eight exposure models outlined in Section 2.3 across the four drug-ADR pairs examined in the case study. The models are arranged from the best (based on the BIC) on the left to the worst on the right. In three of the four cases, the BIC value for the null model far exceeds the range of y-axis. The rounded BIC values are added to the respective bars.

The results for penicillin and anaphylaxis, presented in the upper left corner, align with the anticipated model: current use. The BIC score, and consequently the posterior probability of the null model, strongly indicate the presence of an association. As mentioned earlier,

although the decaying model may be more appropriate, the time resolution in quarter years is inadequate for distinguishing between this model and the current use model.

For DOACs and GI bleeding, the past use model attains the best BIC score. Analogous to the previous drug-ADR pair, the BIC score for the null model substantially exceeds the scores of the other models, suggesting a strong association.

A similar pattern emerges for antipsychotics and type 2 diabetes, with the past use model yielding the best fit, closely followed by the current use model. Once more, the inadequacy of the null model's fit implies an association between the drug and ADR.

For the negative control, antibiotics and GI bleeding, the past use model performs best as well. Despite the null model having the least favorable performance, its BIC score is comparable to the scores of the other models. In contrast, for the other three drug-ADR pairs, the difference between the null model's BIC score and the BIC scores of the other models was much larger.

To investigate why the past use model is clearly preferred in three out of the four drug-ADR pairs under consideration, we delve deeper into this preference and examine the BIC values for the past use exposure model across all values of the parameter p. This exposure model represents a scenario where the patient is at maximal risk during and for an extended period after dispensation, where the length of the period equals p. See equation (6) for the definition and Figures 2k and 2l for examples. The parameter p ranges from 1 to $T - 1$ (equaling 55 in our case). Figure 7 presents the BIC values for both the current use model and past use models across all permissible values of p. The current use model is represented in orange and positioned at $p = 0$ since the past use model is, in that case, equivalent to the current use model.

For DOACs and GI bleeding, the BIC score reaches its minimum at $p = 3$, suggesting that the risk of experiencing GI bleeding remains elevated for approximately 3 quarters after the last dispensation. A similar trend is noted for antipsychotics and type 2 diabetes, with the lowest BIC value occurring at $p = 4$, corresponding to one year. In other words, the risk of being diagnosed with type 2 diabetes stays heightened and decreases after one year following the last dispensation.

The change in BIC scores with the parameter p for the negative control, antibiotics and GI bleeding, reveals an interesting pattern that may partially explain why the past use model is favored in a scenario where the drug and the ADR considered are unrelated. The minimum occurs around $p = 26$, corresponding to 6 or 7 years after the last dispensation. It is highly improbable that exposure to antibiotics increases the risk of GI bleeding so many years later. The method likely detects the natural increase in risk with age, rather than indicating a genuine relationship between antibiotics and the ADR GI bleeding.

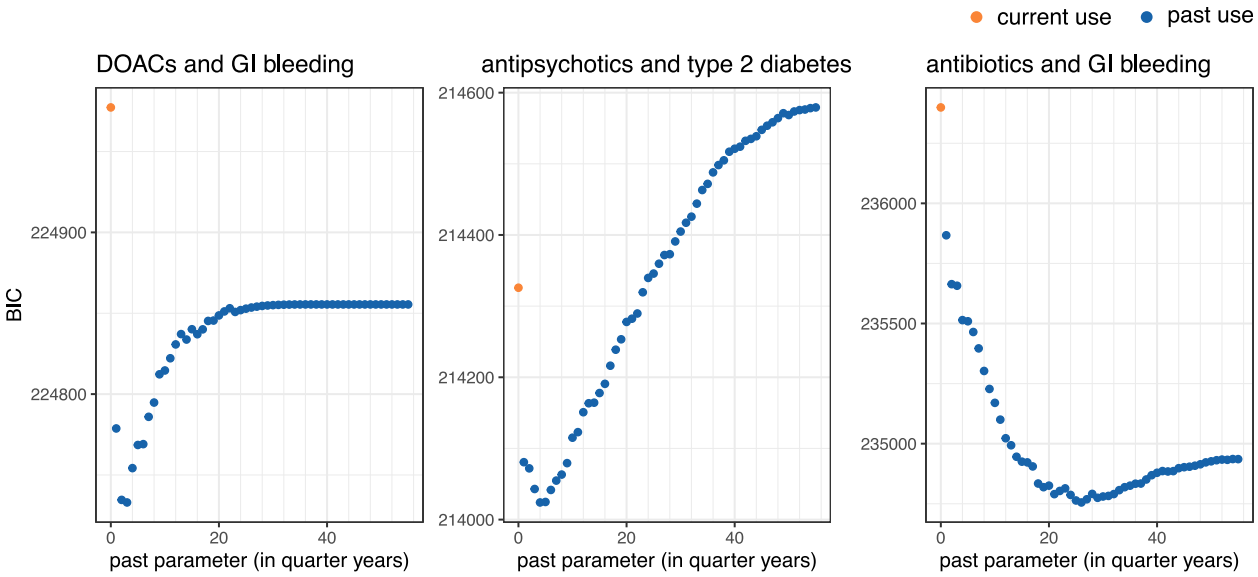

Figure 7: The BIC values for the current use and past use models for three drug-ADR pairs, for all possible parameter values of p (corresponding to quarter years). The value of the current use model is depicted in orange and positioned at $p = 0$, as the past use model is equivalent to the current use model for that specific parameter value.

8 Conclusions and Discussion

In this paper, we introduced a versatile exposure model framework designed to capture various longitudinal relationships that can occur between a drug and an ADR. The framework allows for the estimation of model parameters based on EHC data using maximum likelihood. We suggest the utilization of the BIC to select the most suitable model. Furthermore, the BIC has a direct connection with the models' posterior probabilities, see Section 2.5. This feature makes the approach applicable in a pharmacovigilance context, where the posterior probability of the null model, representing no association, can be used to decide whether a drug-ADR pair yields a signal. Additionally, it facilitates the use of Bayesian false discovery rate procedures [46] to create a shortlist for the committee of medical experts. A main feature of the proposed framework is that it cannot only aid in signal detection, but also allows for exploring the nature of the relationship between the drug and ADR.

We explored the effectiveness of this approach through a simulation and case study. For the simulation study, we developed a unique EHC data simulator capable of simulating any exposure model.

The simulation study demonstrates the capacity to, under certain conditions, determine the presence of an association between a drug and ADR, and accurately identify the correct exposure models. Key factors influencing the performance include: 1) the number of patients that were exposed to the drug (μ_E), 2) the probability of experiencing the ADR when the

patient is at maximal risk (π_1), and 3) the disparity between the probabilities of experiencing the ADR when the patient is at maximal and minimal risk ($|\pi_1 - \pi_0|$). Detecting signals becomes challenging when both the number of exposed patients (approximately 1%) and the frequency with which the ADR occurs are low. Performance improves with a higher number of exposed patients and/or increased ADR frequency.

The performance in confirming an association and identifying the correct model improves when there is a larger difference between the probabilities of the ADR occurring when at minimal and maximal risk. However, this principle does not hold for the withdrawal model. Exposure models reflecting delayed effects are frequently misidentified as the past use model, see Section 7.1. This misattribution can occur since, for such models, the ADR risk increases when the exposure starts and remains high until after exposure, not unlike the trend modeled by the past use model (see Figure 2). Correctly identifying the long-term model poses a significant challenge under almost all of the considered simulation settings. For an interactive exploration of all simulation study results, visit `https://exposuremodels.bips.eu`.

In the case study outlined in Sections 6 and 7.2, we applied the exposure model method to four drug-ADR pairs where the true temporal relationships are, at least approximately, known. Utilizing a data set consisting of insurants from two German SHIs, totaling over 1.2 million individuals, we successfully identified penicillin and anaphylaxis, where the current use model is chosen based on the BIC/posterior probability. For the other three drug-ADR pairs, the past use model was selected. While this is partly expected for the pairs DOACs and GI bleeding, and antipsychotics and type 2 diabetes, it was surprising for the negative control, antibiotics and GI bleeding.

When examining the BIC scores associated with the parameter values of the past use model, see Figure 7, we found that for the drug-ADR pairs DOACs and GI bleeding, as well as antipsychotics and type 2 diabetes, the optimal value tends to center around one year after the last dispensation. In contrast, for the negative control, the optimal value lies around 6 to 7 years. The preference for the past use model in the negative control case appears to be driven by the natural increase in the risk of GI bleeding with age, rather than indicating a genuine relationship between the drug and the ADR. The assumption of the exposure model that the baseline risk (π_0) remains constant over time may lead to misclassification. It would be interesting to explore how the age of an individual could be included to address this issue.

Prior efforts to employ exposure models were undertaken by Van Gaalen et al. [54, 55]. However, their approach considers a limited number of exposure models, and the process of estimating model parameters based on the available data is not clearly defined. Furthermore, the application of their method in a pharmacovigilance context is unclear as well.

One aspect to consider in the current study is that it does not include a comparison with other signal detection methods available in the literature[13]. To undertake such a com-

parison, a significant expansion of the simulation set-up is required. This expansion would involve simulating multiple drug-ADR pairs with varying exposure models simultaneously. Moreover, it is important to take into account different types of thresholds used to define a signal, as these can vary considerably from one method to another [11]. We plan to explore this comparison in future research.

The formalization of EHC data, see Section 2.1, disregards differences in dosage [54] and only captures whether the patient was exposed or not. One could account for dosages by defining the drug history as a real-valued random vector, i.e., $\boldsymbol{X}^k \in \mathbb{R}_+^{T_k}$, rather than a binary one. The reason why we opt for a binary representation is that other signal detection methods do not account for dosage as well, with notable exception of the work by Van Gaalen et al.[54].

As discussed in Section 2.2, we assume that the occurrences of an ADR at different time points are independent given the drug exposure history. However, this assumption may be particularly strong for certain types of ADRs, such as anaphylaxis and myocardial infarction, where a patient is unlikely to be treated with the same drug again after experiencing such a reaction. One potential extension to the model is to incorporate not only the drug history but also the ADR history. This extension involves modeling the conditional probability distribution $\mathbb{P}(\boldsymbol{Y}(t) \mid \boldsymbol{X}(1:t), \boldsymbol{Y}(1:t-1))$ rather than just $\mathbb{P}(\boldsymbol{Y}(t) \mid \boldsymbol{X}(1:t))$. While our framework could accommodate this extension, it would drastically increase its complexity. Nevertheless, it could prove to be a valuable avenue for future research.

Furthermore, it would be intriguing to explore the influence of model misspecification, particularly in scenarios where the true simulated model deviates from the predefined set of models considered by the method. Investigating such scenarios can provide insights into the method's ability to detect associations between a drug and ADR, even when temporal relationships differ from those explicitly considered. Encouragingly, the results of the case study suggest that, to some extent, the correct exposure model can still be identified even in the presence of disparities between the true and selectable models.

In theory, there is no restriction on the number of exposure models that can be simultaneously considered. Nevertheless, it is important to take into account that as the number of exposure models increases, so does the likelihood of selecting one of them over the null model. As a result, the likelihood of generating a signal for a drug and ADR increases with the number of exposure models. One potential approach to address this is by applying a false discovery rate correction to the models for each drug-ADR pair individually. However, the challenge lies in determining how to incorporate this correction alongside a false discovery rate control procedure for all drug-ADR pairs when creating a shortlist.

Employing a Bayesian approach for estimating exposure model parameters provides the advantage of incorporating prior knowledge into the modeling process. This is particularly

valuable when existing knowledge is available, e.g., the probabilities of experiencing the ADR at maximal and minimal risk are close to zero. Similarly, applying a Bayesian prior to the exposure models themselves enables consideration of how frequently a specific temporal relationship is expected to occur. However, it is challenging to select an appropriate prior, given the difference in power to detect various models, as seen in the simulation study, see Section 7.1.

Exploring alternative methods for model selection beyond the BIC could be worthwhile. Even though the BIC has the advantage of being related to the posterior probability, it might be beneficial to consider other selection techniques, particularly since the BIC appears to be excessively conservative.

Author contributions

The contributions are organized according to the Contributor Roles Taxonomy (CRediT), see `https://credit.niso.org/`.

LD: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review & editing. **TS:** conceptualization, resources, investigation, validation, writing – review & editing. **RF:** conceptualization, data curation, funding acquisition, investigation, project administration, resources, supervision, writing – original draft, writing – review & editing.

Acknowledgments

We gratefully acknowledge Oliver Scholle for his insightful contributions to the application study and the interpretation of the results. The authors would also like to thank the statutory health insurances, namely hkk Krankenkasse and AOK Bremen/Bremerhaven, for providing the data used in this study.

Financial disclosure

This work was supported by the innovation fund ('Innovationsfonds') of the Federal Joint Committee in Germany (grant number: 01VSF16020).

Conflict of interest

LD, TS and RF are working at an independent, non-profit research institute, the Leibniz Institute for Prevention Research and Epidemiology – BIPS. Unrelated to this study, BIPS

occasionally conducts studies financed by the pharmaceutical industry. Almost exclusively, these are post-approval safety studies (PASS) requested by health authorities. The design and conduct of these studies as well as the interpretation and publication are not influenced by the pharmaceutical industry. The study presented was not funded by the pharmaceutical industry and was performed in line with the ENCePP Code of Conduct.

References

[1] ViGiBase description of the World Health Organization (WHO). `https://who-umc.org/vigibase/`. Accessed: 2023-12-03.

[2] R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[3] M. Alomar, S. Palaian, and M. M. Al-Tabakha. Pharmacovigilance in perspective: Drug withdrawals, data mining and policy implications. *F1000Res*, 8:2109, 2019.

[4] J. Arnaiz, X. Carné, N. Riba, C. Codina, J. Ribas, and A. Trilla. The use of evidence in pharmacovigilance: Case reports as the referencesSource for drug withdrawals. *Eur J Clin Pharmacol*, 57:89–91, 2001.

[5] M. Avriel. *Nonlinear programming: Analysis and methods*. Courier Corporation, 2003.

[6] C. Bailey, D. Peddie, M. E. Wickham, K. Badke, S. S. Small, M. M. Doyle-Waters, E. Balka, and C. M. Hohl. Adverse drug event reporting systems: A systematic review. *British Journal of Clinical Pharmacology*, 82(1):17–29, 2016.

[7] E. P. Borrelli, E. Y. Lee, A. M. Descoteaux, S. J. Kogut, and A. R. Caffrey. Stevens-Johnson syndrome and toxic epidermal necrolysis with antiepileptic drugs: An analysis of the US Food and Drug Administration adverse event reporting system. *Epilepsia*, 59 (12):2318–2324, 2018.

[8] D. S. Budnitz, M. C. Lovegrove, N. Shehab, and C. L. Richards. Emergency hospitalizations for adverse drug events in older Americans. *N Engl J Med*, 365(21):2002–2012, 2011.

[9] P. M. Coloma, M. J. Schuemie, G. Trifiro, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, G. Corrao, L. Pedersen, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: The EU-ADR Project. *Pharmacoepidemiol Drug Saf*, 20(1):1–11, 2011.

[10] A. Coste, A. Wong, M. P. Bokern, A. Bate, and I. Douglas. Methods for drug safety signal detection using routinely collected observational electronic health care data: A systematic review. *Pharmacoepidemiol Drug Saf*, 32:28–43, 2022.

[11] G. Deshpande, V. Gogolak, and S. W. Smith. Data mining in drug safety: Review of published threshold criteria for defining signals of disproportionate reporting. *Pharma Med*, 24:37–43, 2010.

[12] L. Dijkstra, M. Garling, R. Foraita, and I. Pigeot. Adverse drug reaction or innocent bystander? A systematic comparison of statistical discovery methods for spontaneous reporting systems. *Pharmacoepidemiol Drug Saf*, 29(4):396–403, 2020.

[13] L. Dijkstra, T. Schink, R. Linder, M. Schwaninger, I. Pigeot, M. Wright, and R. Foraita. A discovery and verification approach for pharmacovigilance using electronic health care data. *medRxiv*, 2022. doi: 10.1101/2022.05.10.22274885. URL `https://www.medrxiv.org/content/early/2022/05/10/2022.05.10.22274885`.

[14] J. Ding, V. Tarokh, and Y. Yang. Model selection techniques: An overview. *IEEE Signal Process Mag*, 35(6):16–34, 2018.

[15] W. G. Dixon, M. Abrahamowicz, M.-E. Beauchamp, D. W. Ray, S. Bernatsky, S. Suissa, and M.-P. Sylvestre. Immediate and delayed impact of oral glucocorticoid therapy on risk of serious infection in older patients with rheumatoid arthritis: A nested case–control analysis. *Ann Rheum Dis*, 71(7):1128–1133, 2012.

[16] W. DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Am Stat*, 53(3):177–190, 1999.

[17] S. Feldman, W. Ammar, K. Lo, E. Trepman, M. van Zuylen, and O. Etzioni. Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA Netw Open*, 2, 2019.

[18] D. Finney. The design and logic of a monitor of drug use. *J Chronic Dis*, 18(1):77–98, 1965.

[19] J. L. M. Garbayo, I. P. Castelló, M. T. F. Soler, M. P. Ribis, et al. Hospital admissions for bleeding events associated with treatment with Apixaban, Dabigatran and Rivaroxaban. *Eur J Hosp Pharm*, 26(2):106–112, 2019.

[20] A. Grosso, I. Douglas, R. Macallister, I. Petersen, L. Smeeth, and A. Hingorani. Use of the self-controlled case series method in drug safety assessment. *Expert Opin Drug Saf*, 10:337–340, 2011.

[21] R. Harpaz, K. Haerian, H. S. Chase, and C. Friedman. Mining electronic health records for adverse drug effects using regression based methods. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 100–107, 2010.

[22] R. Harpaz, W. DuMouchel, N. H. Shah, D. Madigan, P. Ryan, and C. Friedman. Novel data mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther*, 91(6):1010–1021, 2012.

[23] U. Haug and T. Schink. German pharmacoepidemiological research database (GePaRD). In *Databases for Pharmacoepidemiological Research*, pages 119–124. Springer International Publishing, 2021.

[24] L. Hazell and S. A. Shakir. Under-reporting of adverse drug reactions. *Drug Saf*, 29(5): 385–396, 2006.

[25] R. I. Holt and A. J. Mitchell. Diabetes mellitus and severe mental illness: Mechanisms and clinical implications. *Nat Rev Endocrinol*, 11(2):79–89, 2015.

[26] I. Karlsson, J. Zhao, L. Asker, and H. Boström. Predicting adverse drug events by analyzing electronic patient records. In *Artificial Intelligence in Medicine: 14th Conference on Artificial Intelligence in Medicine, AIME 2013, Murcia, Spain, May 29–June 1, 2013. Proceedings 14*, pages 125–129. Springer, 2013.

[27] T.-L. Kelly, A. Salter, and N. L. Pratt. The weighted cumulative exposure method and its application to pharmacoepidemiology: A narrative review. *Pharmacoepidemiol Drug Saf*, 2023.

[28] T. R. Kosten and L. E. Baxter. Effective management of opioid withdrawal symptoms: A gateway to opioid dependence treatment. *Am J Addict*, 28(2):55–62, 2019.

[29] F. Margraff and D. Bertram. Adverse drug reaction reporting by patients: An overview of fifty countries. *Drug Saf*, 37:409–419, 2014.

[30] Y. Moride, F. Haramburu, A. A. Requejo, and B. Begaud. Under-reporting of adverse drug reactions in general practice. *Br J Clin Pharmacol*, 43(2):177–181, 1997.

[31] Z. Mosenifar. Population issues in clinical trials. *Proc Am Thorac Soc*, 4(2):185–187, 2007.

[32] R. E. Murray, P. B. Ryan, and S. J. Reisinger. Design and validation of a data simulation model for longitudinal healthcare data. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1176. American Medical Informatics Association, 2011.

[33] J. A. Nelder and R. Mead. A simplex method for function minimization. *Comput J*, 7 (4):308–313, 1965.

[34] A. I. Neugut, A. T. Ghatak, and R. L. Miller. Anaphylaxis in the United States: An investigation into its epidemiology. *Arch Intern Med*, 161(1):15–21, 2001.

[35] G. N. Norén, T. Bergvall, P. B. Ryan, K. Juhlin, M. J. Schuemie, and D. Madigan. Empirical Performance of the Calibrated Self-Controlled Cohort Analysis Within Temporal Pattern Discovery: Lessons for Developing a Risk Identification and Analysis System. *Drug safety*, 36:107–121, 2013.

[36] I. J. Onakpoya, C. J. Heneghan, and J. K. Aronson. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: A systematic review of the world literature. *BMC Med*, 14(1):1–11, 2016.

[37] P. Perucca and F. G. Gilliam. Adverse effects of antiepileptic drugs. *Lancet Neurol*, 11 (9):792–802, 2012.

[38] J. Reps, J. Feyereisl, J. M. Garibaldi, U. Aickelin, J. E. Gibson, and R. B. Hubbard. Investigating the detection of adverse drug events in a UK general practice electronic health-care database. *arXiv preprint arXiv:1307.1078*, 2013.

[39] P. Routledge. 150 years of pharmacovigilance. *Lancet*, 351(9110):1200–1201, 1998.

[40] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10 (3), 2015.

[41] D. Schroeder. Statistics: detecting a rare adverse drug reaction using spontaneous reports. *Reg Anesth Pain Med*, 23(6):183–189, 1998.

[42] M. J. Schuemie. Methods for Drug Safety Signal Detection in Longitudinal Observational Databases: LGPS and LEOPARD. *Pharmacoepidemiology and Drug Safety*, 20 (3):292–299, 2011.

[43] G. Schwarz. Estimating the dimension of a model. *Ann Stat*, pages 461–464, 1978.

[44] H. Shin, S. Park, and S. Lee. Adverse drug reaction analysis methods and research trends by data sources for post-marketing surveillance. *J Health Info Stat*, 2022.

[45] S. Singh and Y. Loke. Drug safety assessment in clinical trials: Methodological challenges and opportunities. *Trials*, 13:138 – 138, 2012.

[46] J. D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat*, 31(6):2013–2035, 2003.

[47] M. Suling and I. Pigeot. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics*, 4(4):607–640, 2012.

[48] The Observational Medical Outcomes Partnership (OMOP). Research partnership announces competition winners: OMOP Cup challenged contestants to develop algorithms to improve drug safety. `https://fnih.org/press-release/research-partnership-announces-competition-winners/`, May 2010.

[49] G. Trifiro, V. Patadia, M. J. Schuemie, P. M. Coloma, R. Gini, R. Herings, J. Hippisley-Cox, G. Mazzaglia, C. Giaquinto, L. Scotti, L. Pedersen, P. Avillach, M. C. J. M. Sturkenboom, J. van der Lei, and EU-ADR Group. EU-ADR healthcare database network vs. spontaneous reporting system database: Preliminary comparison of signal detection. *Stud Health Technol Inform*, 166:25–30, 2011.

[50] P. Tubert, B. Bégaud, F. Haramburu, and J.-C. Péré. Spontaneous Reporting: How many Cases Are Required to Trigger a Warning? *British Journal of Clinical Pharmacology*, 32(4):407–408, 1991.

[51] P. Tubert, B. Bégaud, J.-C. Péré, F. Haramburu, and J. Lellouch. Power and weakness of spontaneous reporting: A probabilistic approach. *J Clin Epidemiol*, 45(3):283–286, 1992.

[52] P. Tubert-Bitter, B. Bégaud, and I. Ahmed. Comparison of two drug safety signals in a pharmacovigilance data mining framework. *Stat Methods Med Res*, 25:615 – 629, 2016.

[53] P. G. M. van der Heijden, E. P. van Puijenbroek, S. van Buuren, and J. W. van der Hofstede. On the assessment of adverse drug reactions from spontaneous reporting systems: The influence of under-reporting on odds ratios. *Stat Med*, 21(14):2027–2044, 2002.

[54] R. D. van Gaalen, M. Abrahamowicz, and D. L. Buckeridge. The impact of exposure model misspecification on signal detection in prospective pharmacovigilance. *Pharmacoepidemiol Drug Saf*, 24(5):456–467, 2015.

[55] R. D. van Gaalen, M. Abrahamowicz, and D. L. Buckeridge. Using multiple pharmacovigilance models improves the timeliness of signal detection in simulated prospective surveillance. *Drug Saf*, 40:1119–1129, 2017.

[56] A. van Hylckama Vlieg, F. Helmerhorst, J. Vandenbroucke, C. J. M. Doggen, and F. Rosendaal. The venous thrombotic risk of oral contraceptives, effects of oestrogen dose and progestogen type: Results of the MEGA case-control study. *BMJ*, 339, 2009.

[57] T. van Staa, H. Leufkens, L. Abenhaim, B. Zhang, and C. Cooper. Use of Oral Corticosteroids and Risk of Fractures. *Journal of Bone and Mineral Research*, 15(6):993–1000, 2000.

[58] P. Waller and M. Harrison-Woolrych. *An introduction to pharmacovigilance*. John Wiley & Sons, 2017.

[59] D. Weintraub, R. Buchsbaum, S. Resor Jr, and L. Hirsch. Psychiatric and behavioral side effects of the newer antiepileptic drugs in adults with epilepsy. *Epilepsy Behav*, 10 (1):105–110, 2007.

[60] World Health Organization. The importance of pharmacovigilance. Technical report, 2002.

[61] G. Zaccara, D. Franciotta, and E. Perucca. Idiosyncratic adverse reactions to antiepileptic drugs. *Epilepsia*, 48(7):1223–1244, 2007.

[62] I. Zorych, D. Madigan, P. Ryan, and A. Bate. Disproportionality methods for pharmacovigilance in longitudinal observational databases. *Stat Methods Med Res*, 22(1): 39–56, 2013.