

Studies of Applications of Expected Information Theory to Overcome Current Problems in Adverse Drug Reaction Analytics

Barry Robson¹ and Jim Weisman¹

¹Inginer Inc., Cleveland, Ohio, USA, and The Dirac Foundation, Oxfordshire, UK

Article Type	Original Research
Volume / Issue	1 / 1:6 (2026)
DOI / ISSN	https://doi.org/10.5281/zenodo.22283078 / ISSN: 3144-0665
Correspondence	Barry Robson, PhD DSc Email: barryrobson@inginer.com Inginer Inc., 11000 Cedar Ave Ste 100, Cleveland, Ohio
Received / First Decision / Accepted	July 27, 2026 / August 19, 2026 / September 2, 2026
Disclosure	Funding: None. Conflicts of Interest: None.

Abstract

Purpose: Current adverse drug reaction (ADR) signaling methods used in pharmacovigilance have an extensively empirical nature. It is here described how a signal is developed based on the Theory of Expected Information (TEI), in which partially summated zeta functions emerge naturally from integrals of information measures over prior probability densities. This also provides a convenient and better basis for analyzing, comparing, and even combining, other methods, so that this is to some extent also a review and position paper.

Methods: Glass Box AI-style methods, akin to high-dimensional data mining and prediction from it, were used.

Results: Examples show how hundreds of thousands of reusable knowledge elements obtained from analysis of the FAERs drug reports database contain statistical information that can be inspected for insight and collectively analyzed and used for alerts, as well as for analytical inference and prediction. These elements can be obtained even for single events, and in principle even for joint events that should have been seen, based on prior expectation, but were not seen.

Conclusions: The above is shown to represent a more formal approach to ADR Analytics that is smoothly extensible from sparse data. Sparse data appear in (a) early warning signals in a single year’s quarter, and (b) in more detailed analyses involving joined patient demographic, clinical, and other data to assist in stratification, segmentation, and drug interaction warnings.

Keywords: Adverse Drug Reaction | Reporting | Analytics | Stratification | Segmentation

Key Points

- Current methods in adverse drug reaction analytics are complex and tend to contain empirical parameters including those used to unhighlight alerts when data is sparse. While “sparse” depends on authors, it generally means less than about 10-20 observations. Methods are usually described as failing when data about a new drug is very sparse, typically meaning 5 or less observations.
- A Theory of Expected Information based on the emergence of the partially summated zeta function from integrals, that has proven successful in other areas such as retrospective cohort studies and clinical trial selection, is adapted to overcome these difficulties.
- Some current methods and formulas are rewritten using the Theory and compared.
- A Glass Box prediction method based on the above theory typically generates thousands to millions of probabilities as knowledge elements or “tags” that combine to make predictions, notably including as Likelihood Ratios. It may thus be considered a “Large Probability Model”. Each “tag” carries useful reusable information and statistics.
- The FDA FAERS ADR report and DEMO demographic datasets are mined, with the emphasis on the last 18 million entries to establish early warning reports, and the same principles are then applied to mine clinical health records in a preliminary study to illustrate joining to more detailed patient information.

Plain Language Summary

Current methods for monitoring bad responses of patients to drugs (e.g., allergies, adverse reactions) are argued to be somewhat empirical in nature and not suited to the sparse data available for new drugs on the market. Here, we describe methods based on a Theory of Expected Information that emerges naturally from the basic mathematics of information. Examples show how hundreds of thousands of reusable knowledge elements obtained from the analysis of a drug reports database contain statistical information that can be inspected for insight, alerts, and predictions, even for single patient events (cases). This is argued to represent a more formal approach to adverse drug reaction analytics. It is also briefly discussed as to how information about adverse reactions might be joined with the richer detail in patient records or deduced from them; this will be reported in detail elsewhere.

1 Introduction

1.1 Current Problems in Adverse Drug Reactions

To enable pharmacovigilance, adverse drug reaction (ADR) report databases were implemented in the 1960s, partly motivated by the Thalidomide tragedy. Up to the 1990s, signal detection was mostly manual and largely based on simple traditional statistical techniques or just clinical judgment. The US Food and Drug Administration (FDA) and similar bodies elsewhere favored, for almost all purposes, traditional frequentist statistical methods like those familiar from High School [1] and, later, Bayesian methods [2]. For ADRs the FDA etc. have approved of datamining and related data-analytic methods applied to reporting databases such as the FDA's Adverse Event Reporting System (FAERS but often referred to as the FAER database) [3] for marketed drugs. It has been an important public health pursuit for researchers for several years. The methods used are primarily of the type of study called a disproportionality analysis which from the earliest days of ADR report databases has been used to detect side effects to drugs following their approval (e.g. Ref [4]).

In recent years there has been a flood of submissions to journals applying such methods to ADR report databases, many claiming causality between the drug in question and the ADR reported, and many of these have suffered from methodological difficulties [5]. Since 2023-2024, this has reached a sufficient level of concern for journals that they have discouraged submission (e.g. Ref [6]). The term “disproportionality analysis” used above relates to the occurrence of joint events (two or more in concurrence) more than might have been expected on some kind of prior basis, e.g., associating by chance. On first consideration, that would not seem to be a problem. Association measures, odds, and odds ratios of various types, or their logarithms, are all used to detect ADRs and belong to, or closely relate to, that disproportionality class, and have of course been extensively and successfully used in many other areas of data analysis. They are essentially information measures and, as the difference of what was expected and what is observed, would seem well suited to act as warning signals about drug safety. However, as used, the methods are not well suited to data like FAERS [5]. They are not suited to high-dimensional and/or sparse data, or the nature of the entries on the database including, for example, records that contain duplications and corrections regarding the same ADR event, as discussed in the present paper. About 40% of the records attributing the ADR to a drug can be duplicates in FAERS. Entries are extensively subjective and there is also an absence of contrary evidence as discussed in Section 1.3 and throughout this paper. It also appears to be too easy for many current efforts to pick out ADR signals from reports databases whose ‘causality’ will often be shown on follow-up analysis to be false positives, causing unnecessary alarm [5].

These limitations in the structure, content, and reliability of ADR report databases are recognized at national levels. The US FDA’s SENTINEL project is underway to include more patient demographic and clinical data [6]. However, this is still work in progress, data is not widely available, and much content comes from claims/insurance data lacking laboratory results and other clinical details. At the time of writing, efforts to include Electronic Health Records (EHR) are still in very early stages, and in the US managed healthcare system, will likely still be partially fragmented due to lack of Continuity of Care (COC) [7]. Ideally, ADR entries should be a natural intrinsic feature of patient’s cradle-to-grave Longitudinal Health Records (LHRs) implemented in a structured manner enabling efficient analysis. This was the motivation of a project [8] for the South Korean Electronics and Telecommunications Research Institute, providing an ontology with which the methods described here will be compatible (see Methods Section 3.1). But as the optimal data source for ADR detection is still in flux, any detection method should be compatible with any data set, with a focus on extracting information from the data in the least empirical and most transparent, explainable, and verifiable way possible.

1.2 A New Kind of Analysis is Needed

A new kind of analysis is critically needed because the patients, administrative authorities, and the pharmaceutical industry depend on reasonably reliable warnings and a rational basis for speedy and judicious response. An example of a recent effort towards this aim is the Markov Chain Signal Generation (MCSG) method [9]. Although this is promising based on the results, it depends on constructing a hierarchical Bayesian model for background rates, then background rate samples are drawn from the model parameters using an iterative Markov Chain Monte Carlo method, and at each iteration, any counts whose probability is low given current background rate estimation are removed from the computation that samples the next set of background rates. This complexity is typical of current efforts and in the associated paper it appears that the performance of the algorithm could be sensitive to changes in data, compared with a single formula based on counts. There are also limitations of Markov Chain Monte Carlo methods as regards independence assumptions. Be

that as it may, it is a more fundamental information-theoretic approach that is explored here. More traditional contenders that seemed promising (except for the important caveats below) have been Black Swan Theory [10] and Dismal Theory [11]. The 21st Century has seen an increasing interest in the statistics and risk/benefit analysis of “Black Swan Events” (BSEs) mostly referring to large-scale catastrophes that are rare and not easily predicted. The above two theories address those but tackle risk assessment differently. Drug approval or marketing suspensions seem to fit the label “catastrophe” because they often cost more than many natural disasters and may deprive some patients of their only means of treatment [12-15]. Indeed, much of the language of BSEs is often used by organizations such as the European Medicines Agency and the WHO-Uppsala Monitoring Center that monitor adverse drug reactions (ADRs) in post-marketed drugs. However, BSE methods have not been adopted by the FDA etc. Unlike a major disaster as a BSE, the occurrence of each individual ADR is not always obvious, and reports are subjective and sometimes in error as to true cause of symptoms. Also, unlike BSEs, ADRs are neither rare nor unexpected, although the timing and nature of them is not usually known in advance. Some 90% of drugs fail in clinical trials [15] and increasingly demand special statistical [15-18] as well as ethical [19] and personalized considerations, e.g., by including patient genomics [20]. The most challenging arena for a new drug is after its FDA approval and product launch when pharmacovigilance [20,21] must address a much larger kind of ongoing “trial”. Some 131 million Americans take over 4 billion prescriptions annually, resulting in about 81 million adverse reactions, 2.7 million hospitalizations, and over 100,000 fatalities. Such large populations and human diversity make it inevitable that ADRs will occur, and some 20% of newly marketed drugs require withdrawal or warnings [19].

1.3 Major Challenges and Purposes of this Paper

This paper relates to the discipline that may be dubbed ADR Analytics, a field in which Refs [4,21-24], by other authors, provide examples. The proposed approach uses well-based information theory which is not confined to preliminary signaling alerts in Post-Marketing Surveillance (PMS). Consequently, it is smoothly extensible to pharmacoepidemiology, Comparative Effectiveness Research (CER) that evaluates how a drug performs relative to other available treatments in real-world conditions, Health Economics and Outcomes Research (HEOR) that evaluates the economic, clinical, and humanistic outcomes of a drug, repurposing and pricing, and Clinical Decision Support (CDS). The present authors have already helped show that the information-theoretic approach used here is relevant to these areas (e.g. Ref [25]). The proposed approach also explores how initial signals could be gleaned from better quality Longitudinal Health Record (LHR) data. It refers to some of the ideas used in BSE theory such as information content of sparse versus plentiful contributions to data, but it is not BSE theory per se. The proposed methodology gives a more accurate signal (at fewer occurrences) than prominent standard methods though further studies with a more extensive list of drugs is required. The focus is on illustrating how a primarily information-theoretic approach can enable a deep investigation and comparison of methods. At present, it allows for initial inferential analysis with linked FAERS DRUG and DEMO datasets.

The main ADR reporting databases (RDs) have presented challenges even since developments in current ADR signal detection (e.g. Refs [21-24]). Many widely used ADR-RDs, such as those of the FDA [1] and WHO [23], are ‘messy’, requiring considerable processing that can vary between researchers (see below). Only patients who have suspected ADRs appear on the data base so that the data is biased: counterfactuals, such the event suspected to be an ADR could potentially happen without the named drug, are missing. The data fall short of representing the “four pillars of evidence” of epidemiology [25]. Patients, caregivers, relatives, physicians, and pharmaceutical companies are allowed to make reports but have varied expertise and inevitable different

perspectives and biases. Records for the same event for a specific patient can appear several times, most repeats being more detailed follow-ups or corrections.

The important data to look at for a newly marked drug are inevitably sparse. Early warning indications that a newly marketed drug could risk suspension are crucial and allow little time for pharmaceutical companies to avoid or reduce risk of drug suspension by appropriate notifications and contraindications on labeling, and by stratification and segmentation. An informal rule of thumb for considering possible suspension of a new drug is to use the first three (possibly four) well-documented ADR cases (see Section 1.5). This is naturally impacted by the rate of appearance of new ADR data, and some statistics on new data releases and monitoring are as follows. The US FDA FAERS database discussed below is updated on a quarter year basis, of about 2 million records per quarter of which about 10,000 are different drugs names. Various estimates state that about 40% may be duplicates. FDA's Center for Drug Evaluation and Research has approved roughly 35–55 novel drugs per year in recent years, averaging about 9–14 per quarter. These products begin to appear in FAERS once adverse event reports are submitted. Recently, daily alerts have been published. Regulatory agencies often require enhanced monitoring (e.g., Risk Evaluation and Mitigation Strategies, mandatory reports) in the first 6 months of a new drug release. Most new safety signals emerge within the first 1–2 years of widespread use and companies may be required to submit Periodic Benefit-Risk Evaluation Reports (PBRERs/PSURs) every 6 months for at least the first 2 years, with the first 18 months being of particular interest.

Reliable information is sparser still. It is helpful that, in the FDA FAERS DRUG dataset, there is not only data that a drug is a primary suspect (PS) but also when a drug is only a secondary suspect (SS) or involved in interactions and possibly not in isolation responsible (I) or simply given concomitantly (C) or shortly prior and not thought responsible. However, these assignments are extensively subjective and can change on later report records for the same case (the “duplicates”). On the FAERS report record, the closest to hard experimental proof of causality is a Dechallenge/Rechallenge test. Symptoms suspected to be an ADR should vanish if an administration of the drug considered responsible is halted and they return if resumed. Unfortunately, as discussed here, these entries are noisy on the FAERS database, and even sparser.

Once an alert signal has occurred, the data that becomes important is sparser still. Cases are distinguished by joining information specific to demographic, clinical, genomic, and other sub-groups of interest [8]. Stratification is usually held to mean dividing data and appropriate populations for prescription into predefined subgroups (strata) based on specific characteristics, e.g. age, sex, geographic region, comorbidities, or concomitant medications, to analyze safety signals within each subgroup. Segmentation refers to using such groups for targeted analysis, communication, or intervention. The terms are used interchangeably below. The FAERS DEMO (demographic dataset) is very limited, essentially confined to sex, age, and weight. Such information is typically all that can be directly joined because there are typically no linkers to further demographic and clinical details of patients.

Some statistics on report data published in latest 3-, 6-, 18- and 24-month blocks are of various interest below, involving mining up to some 16 million report records for studies of ADRs in general and for sifting to discover statistics of ADRs for drugs of specific interest.

1.4 Adverse Drug Reaction Report Databases and Current Methods

A pioneering effort to address some of the above issues was made by Bate in his 2003 PhD Thesis [23], and like the approach described below and used here, this had an information-theoretic

flavor. In his Thesis, Bate developed or refined the Bayesian Confidence Propagation Neural Network (BCPNN). He used the WHO database that contains millions of reports of suspected ADRs from some 70 countries since 1968, maintained by The Uppsala Monitoring Centre (UMC). The UMC was also concerned about detecting signals of a possible causal relationship between possible ADR and a specific drug (again, see Section 1.4). Bate's thesis also refers to the WHO Vigibase that contains some details about the reported reaction, patient age, sex, and medical history, and outcome. Other examples of reporting systems and ADR databases are the Yellow Card Scheme (MHRA) in the UK, EudraVigilance in the EU, and FAERS (FDA Adverse Event Reporting System) in the USA. Bate's basic alert measure is one of disproportionality based on classical association constant K (or odds ratios from it), essentially a modified estimate of $P(\text{Drug-X, ADR}) / [P(\text{Drug-X}) \times P(\text{ADR})]$ where the drug-X means not only a specific drug but the primary suspect for the symptoms suspected by the reporting healthcare professional to be an ADR. In contrast, odds and odds ratios discussed below do include "not" cases. The approach described in Bate's thesis is used by the WHO [24] as the alert signal in the Confidence Propagation Neural Network (BCPNN) method of Bate, Lindquist, Edwards, Olsson, Orre, Lansner, De Freitas, and others. UMC still reports and uses Bate's BCPNN style information content as one of its disproportionality measures, but signal detection evolves: since 2014 the UMC adopted vigiRank as its core statistical signal-detection scoring method, which combines disproportionality (IC-like information) with other evidence aspects (report completeness, recency, geographic spread, narratives) to prioritize signals. Nonetheless, Bate's method remains a foundational part of UMC/WHO analytics. Many comments on the IC/BCPNN approach in the present paper refer to the description in Bate's thesis [23]. Note that the disproportionality metric there does not stand alone but is described as refined by a "neural network", although this is a probabilistic neural model, not a deep neural network. In effect, the network learns statistical associations between them using Bayesian probabilities to learn joint and marginal distributions. Today, arguably, Neural Nets tend not to be considered as the best approach for the above kind of work because of entrapment in locally optimal solutions. Note that the novel approach discussed in the present paper replaces learning by high-dimensional datamining, so local entrapment is avoided (and real probability estimates are visible).

1.5 Previous Work on the Approach Proposed Here

The above major approaches in data analytics rest at least partly on information theory and, as noted above, that is particularly explicit in Bate's approach (his Information Component IC is expressed in bits). Although the method proposed below incorporates very recent developments in high dimensional data mining and construction of prediction networks from it, a core part of its theoretical basis also uses information theory. It focuses, however, on a particular aspect that has an established history outside ADR analytics. Its use is motivated by the comments on sparse new drug data in Section 1.3. They imply the need for a smooth transformation between use of sparse and plentiful data, while the impact of the amount of data on prediction should have a rigorous information-theoretic basis. For example, it should correctly treat the extent to which 3 out of 5 observations carry less contributing information than 300 out of 500. Note that even if data is plentiful, high dimensional data mining can cover complex multifactor joint events for which direct numbers of observations are sparse. This kind of situation also arose in the early days of protein structure science when data was generally sparse, but researchers wanted to consider the impact of local structure of interactions between more and more amino acid residues. The appropriate general kind of method has been developed in bioinformatics over several decades [26-30] and more recently for biomedical analytics and clinical decision support [31-34] and is called the Theory of Expected Information (TEI). The algorithms making use of it also have advantages in avoiding assumptions of independence that are

in other inference methods such as Pearl's Bayes Net [35,36]. The application of TEI is supported by the Q-UEL language (e.g. Refs [7,34,37]). The characteristic feature of the TEI is that integrations to deduce expected values of information measures result in expressions involving partially summated zeta functions [31,32]. These have also emerged as a feature of some Black Swan approaches to adverse events [10,11,38], at least when relating to Zipf's law [39] to handle very rare and costly "one-off" events. Comparisons between accepted approaches to ADR signaling have shown that the different measures concur only when data is not extremely sparse (typically meaning exceeding 20 observations and certainly a 100 or more) and may differ significantly when moderately sparse, say between 5 and 10 observations, with 5 or less being considered as very sparse, e.g. Ref [40]. Even 3 or less severe ADR reports are of concern. Frequentist statistics [41] would consider their applications to sparse ADR data as "underpowered".

2 Theory

2.1 Theory of Expected Information

Pioneering approaches to ADR signaling are ultimately related to quantification of information [42] but not in the way introduced in Section 1.5. The core of TEI is relatively simple. Bayes' equation gives the posterior probability $\Pr(P|D)$ as follows.

$$\Pr(P | D) = \Pr(D | P) \Pr(P) / \Pr(D) \quad (1)$$

$\Pr(P)$ can be considered as the Bayesian prior belief in a "true" underling probability value P . The following was the original form [27,28] for the expectation of the so-called star function $^*(x=1;2; y)$. It is a log odds measure as a potential prediction of $x=1$ as opposed to $x=2$, given counts $n(x,y)$ in data $D(x,y)$ for the posterior probability density $\Pr(P|D)$.

$$\begin{aligned} E[* (x = 1; 2; y) | D(x, y)] &= (1/\Pr(D)) \int_0^1 \ln\left(\frac{P}{1-P}\right) \Pr(D|P) \Pr(P) dP \\ &= k(x, y) \int_0^1 \ln\left(\frac{P}{1-P}\right) P^{n[x=1,y]+v[x=1,y]} (1-P)^{n[x=2,y]+v[x=2,y]} dP \end{aligned} \quad (2)$$

Note that $n[x=1,y] + v[x=1,y]$ was represented by $h(x=1,y)$ in the original representation [28] and similarly for $x=2$. Information $\ln(P/(1-P))$ is today called *Log Predictive Odds* (LPO) which is also of importance in the present paper, although the Log Likelihood Ratio (LLR) is more commonly used here, as discussed shortly below. The above remains the basic and historically first form, and it appears as part of the LLR formula and others. Multiplier $k(x,y)$, is the combinatorial expression for normalization. The rest of the notation above is for the general case, but we introduce already the notation $n[]$ for the actual frequency (number of observations) which comes from the likelihood $\Pr(D | P)$ in Eqn.1 and the virtual frequencies $v[]$ that come from the Bayesian prior degree of belief or knowledge $\Pr(P)$. Subsequently, interest is in $x=1$ as an ADR and $x=2$ its absence, or an ADR not due to the drug of interest. The y becomes the drug but in a richer database could include demographic and clinical features. For brevity, the following will normally be used: $d \equiv \text{Drug-X}$, $a \equiv \text{Drug-X-ADR}$, $d' \equiv \text{Other-Drug}$, and $a' \equiv \text{Suspected-Other Drug-ADR}$ (or more generally other-cause-ADR). Eqn. 2 is an integral that results in an expression involving the incompletely summated zeta function $\zeta(s,n)$ for $s = 1$ [27-32]. See especially Ref [28]. Integration of Eqn. 2 yields the following.

$$\begin{aligned} E[* (x = 1; 2; y) | D(x, y)] \\ = \zeta(s = 1, n[x = 1, y] + v[x = 1, y]) - \zeta(s = 1, n[x = 2, y] + v[x = 2, y]) \end{aligned} \quad (3)$$

Going from this more general historical case with mathematician's notation to an informal one for the typical special case of interest here, the most basic application to ADR analytics appears as follows in Eqn 4. This is the notation that will be mostly used from this point.

$$\begin{aligned} E[* (a: a'; d) | D(a, a', d)] &= \zeta(s = 1, n[a, d] + v[a, d]) - \zeta(s = 1, n[a', d] + v[a', d]) \\ &\equiv \zeta(s = 1, n[d, a] + v[d, a]) - \zeta(s = 1, n[d, a'] + v[d, a']) \end{aligned} \quad (4)$$

Here a is an ADR and a' the absence of that ADR, and d is the name of the drug of interest. Later below, the need to distinguish different types of assessment of the role of the drug in an ADR will require a richer notation, but the following general ideas still apply.

The function $\zeta(\)$ is explained in association with Eqn 5 below, but the reason for specifying the equivalence should first be stated. It is easy to set up formulae in terms of several $\zeta(\)$ functions but there is a matter of the order of a and d expressed as joint events, such as $[a, d]$ versus $[d, a]$, that requires attention. This choice of order is purely notational because the comma in $n[a, d]$ etc. has the force of logical AND. Consequently, $n[a, d] = n[d, a]$, $n[a', d] = n[d, a']$. To relate Eqn. 4 to Eqns. 3 and 2, however, note that the expression to the left of the equivalence sign ' $\equiv$ ' using $[a, d]$ and $[a', d]$ is the form that maps directly to $[x=1, y]$ and $[x=2, y]$ respectively. In contrast, the order $[d, a]$ and $[d, a']$, used to the right of ' $\equiv$ ', is preferred in this paper and probably by most ADR researchers. It is usually a specific drug that comes first to mind for pharmaceutical companies and authorities, especially when the drug is new. Importantly, the ADR researcher is mostly interested in predicting or detecting the adverse drug reaction a as *opposed* to no adverse reaction a' , so naturally $x = 1$ becomes a and $x = 2$ becomes a' , and y represents all the factors on which that is conditional. Note that there will sometimes be further factors or conditions c taken into account, such as age group, so $[d, a]$ and $[d, a']$ should be thought of as readily extensible to $[d, a, c]$ and $[d, a', c]$.

Nonetheless, the importance of x and y in the format of Eqn. 1 is not only because it is the historical, published form and mathematically more rigorous, but also because it is more general: $x=1$ does not have to be the ADR a , and so on. Even in ADR studies, the target to predict, i.e. $x = 1$ versus $x = 2$, in principle could be any of adverse reaction a , drug d , or, as noted above, some further factor or condition c such as patient sex, age group, ethnic group, comorbidities, etc. or a combination of them, and y could in principle usefully be any of the factors then remaining once $x = 1$ and hence $x = 2$ have been assigned. For example, one might wish to predict, or indicate, that age group 50-59 is most susceptible to an adverse effect of the specified drug as opposed to any other age group, given (conditional upon) a specified drug d and an adverse reaction a . The "d before a" order is consistent with the Likelihood Ratio LR that one is frequently interested in calculating, and $LR = P(d|a) / P(d|a')$. Whenever simple odds are used, the LR is more commonly used than the Predictive Odds $PO = P(a|d) / P(a'|d) = P(d, a) / P(d, a')$. It is indeed Eqn. 2 as an estimate of the log PO that historically was tackled first, as noted in comment associated with it: again, the PO is the basis of, and is explicitly or implicitly as part of, most formulae discussed here. As with Eqn. 4, it is always clear from the formulae as to what is being specified relative to what, i.e. what information gain, difference, or change is being measured. This is dramatically important when comparing different alert signal measures.

Unlike use of logs, it is meaningful for $n[d, a] + v[d, a] = 0$. The partially summated zeta function $\zeta(\)$ is, in the broader context, as follows.

$$\zeta(s, n[\] + v[\]) = 1 + 2^{-s} + 3^{-s} + \dots + (n[\] + v[\])^{-s}, \quad s \geq 1 \quad (5)$$

Eqn. 2 gives the case $s=1$. It follows from Eqn. 2 that Eqn. 4 as the expected value of $\ln(P/(1-P)) = \ln(P) - \ln(1-P)$ converges to $\ln(n[d,a]+v[d,a]) - \ln(n[d,a'] + v[d,a'])$ as data increases, with consequences as discussed shortly below. See Ref [32]. This appearance of the zeta function [27,28] was somewhat a surprise because Euler's pioneering work in this area had a different, more general algebraic, purpose focusing on a similar integral over $\log(P)$ and use of his Ψ function, so essentially bypassing ζ , though the results are certainly consistent because of cancellations. Also, the integration in Eqn. 2 initially described $\zeta(s=1, n[\] + v[\])$ as a "number theoretic function" $\#(n[\]-1)$ [28] as also discussed shortly below; the justification for thinking in terms of the fuller zeta function came much later (e.g. Refs [31,32]). Use of $s>1$ relates to higher moments of information. Notably, use of $s > 1$ is also an important parameter of the zeta function for different surprise measures in general, and appears in certain Black Swan theories [37,38]. A surprise measure is one which increases with unusual or early, including meaning sparse, observations. Only $s=1$ is used in this paper, and $n[\] + v[\]$ is usually simply $n[\]$, the number of "observations" (occurrences). It follows from the above that for $s=1$, $\zeta(\)$ relates to natural logarithms of probabilistic measures as information (discussed below,) or their exponentials as disproportionality measures, that are also traditionally used in ADR research. One can use TEI simply by replacing natural logarithmic functions $\ln(\)$. Complex and especially hyperbolic (or split) complex zeta functions are of use (e.g. Ref [46] but beyond present scope.

The virtual frequencies $v[\]$ are analogous to parameters used by current ADR signal methods where they are justified by reference to e.g., β function and Dirichlet distribution parameters, but many numbers added are essentially empirical, adjusted for an initial database and for "unhighlighting" signals not wanted from sparse data where the methods break down. In principle, $v[\]$ can be a real positive or negative number, and in original TEI an absolute prior Dirichlet distribution with all $v[\] = -1$, earlier used throughout [27,28] for reasons explained at the end of Section 2.2. However, for required behavior near $n[\] = 0$ [31], and if there is no justifiable prior information to represent by $v[\]$, then ideally nothing should be added to $n[\]$. More properly stated, consistency with Euler's integral discussed above would require that $n[\] = 1$, because his single parameter analogous to $n[\] + v[\]$ became analogous to $n[\]$ by his use of the β function (consistent with later Dirichlet). In any event, information based on large numbers of observation and small prior belief approaches classical results based on log ratios of counts $n[\]$.

$$E[* (a: a'|D(a, a', d))] \rightarrow \ln(n[d, a]) - \ln(n[d, a']), \quad n[d] \rightarrow \infty \quad (6)$$

Note that $v[\] = 0$ was assumed, meaning no prior belief, and much work below does adopt that position. However, case of importance that may be considered as use of $v[\] \neq 0$ is as follows.

$$v[d, a] = \Pr(a) (n[d, a] + n[d, a']) \quad (7)$$

Then an estimate of expected mutual information $I(a; d)$ is as follows.

$$E[I(d; a) | D(d, a, a')] = \zeta(s=1, n[d, a]) - \zeta(s=1, v[d, a]) \quad (8)$$

Note that, unlike use of log functions, $n[d, a] = 0$ causes no problems and is meaningful: it was expected to see $v[d, a]$ instances, but none were seen. This can be highly significant negative association ($K \ll 1$) and is a very different philosophy from many data mining approach that require "minimum support". The approach requires calculating expected frequencies from significant events "in advance", e.g. as in Eqn. 11 later below, but this is not probative for the reasonably small number of different events encountered in typical ADR analytics. Eqn. 8 here is mainly used as a preliminary tool, for unsupervised data mining to help select targets to predict, and not for many of the numerical

results reported in Results Section 4. However, it is a useful measure alerts to that of Bate (Section 2.3) used for signal alerts for a potential ADR to a specified drug, so for completeness, note the following (for a more formal and detailed explanation see Ref [34] and references therein). First, for any interpretation of number n such as a number of states, then for $n \rightarrow \infty$ we have $\zeta(s=1, n) - \gamma \rightarrow \ln(n)$ where γ is the Euler–Mascheroni constant 0.57721.... For finite n , $\gamma = \zeta(s=1, n) - \ln(n)$ is not really a constant but a function $\gamma(n)$ that quickly convergences to 0.57721.... as n increases. The effect of γ usually almost always cancels out, at least to a good approximation, because the measure of interest is almost always the difference between two information measures defined by zeta functions. Any objection disappears entirely if one takes expected information as *defined* by the zeta function. However, in making use of equations like Eqn. 8, the *expected frequency* for events discussed later below is equal to row sum times column sum divided by total sum, implying the use of *three* zeta functions and in general an odd number. Consequently, when the zeta function is taken as definitional of the above cancellation (by zeta function minus zeta function) that cancellation is then not complete: there is “an isolated” zeta function left over. Eqn. 8 bypasses that problem. Otherwise, if the zeta functions are not considered definitional of information and that role resorts back to the responsibility of natural logarithms, an alternative is to replace each zeta function by a “corrected” form. A simple example that is quick to compute in such cases when v may be real (have a fractional part) including the interpolation is as follows (see Ref [34]).

$$z(v) = \zeta(s=1, v) + (v - \text{int}(v)) \cdot (\zeta(s=1, \text{int}(v+1)) - \zeta(s=1, \text{int}(v))) - \gamma v / (v+1) \quad (9)$$

Here $\text{int}(v)$ is the floor value, the largest integer equal to or less than n . It represents a compromise between zeta and classical bases for information measures in those rare cases where a single zeta function applies. Eqn. 9 is also used for integer values of n below, but makes negligible difference when equations like Eqn. 8 are used. Such simple interpolation methods are not suitable for all kinds of biomedical data. However, the boundary conditions and floor-value operations ($\text{int}(v)$) apply here as defined for sparse count scenarios, as well as for abundant count scenarios, being good estimates for the growth of $\zeta(s, n)$ with n when seen as an ideally continuous smooth, monotonic concave function, and with the important differences $\zeta(s, n_1) - \zeta(s, n_2)$ for $n_1 > n_2$ converging to a plateau as n_1 and n_2 increase. The same algorithm is used for all values of n , with n computed for integer values on startup and stored as lookup functions, except that $\ln(n) + \gamma$ is used for $n > 1000$ in present studies. Greatly increasing the threshold value of n at which this is applied is perfectly feasible on a laptop computer, but increasing that threshold, and representation by look-up tables for a smooth curve with various small increments of n as a real number (rather than integer) gave no perceivable advantages. That we obtain zero information for first observation when using $\ln(1)$ (arguably the most important observation) and $\ln(0) = -\infty$ makes it undesirable as an information measure. While $\ln(1) = 0$ and $\zeta(s=1, 1) = 1$, then $z(1) = 0.7114$, close to 1 bit (precisely, 1.0263... bits). In short, the above method gives plausible values for information given real values lying in between integer value v and $v+1$ for all values of v , using the above linear interpolation based simply on the decimal part of v [34].

2.2 General Principles of Analyzing versus Signaling Adverse Drug Reactions

For present purposes, recall that $s=1$, and $n[d, a]$ is the number of records indicating that drug d is the cause of an ADR a . More generally, however, for analysis, one often wishes to extend this to one or more further conditions c , for example patient demographics and potentially clinical data. The general form mainly used here makes use of the following.

$$n[d, a, c] > 0; v[d, a, c] = 0 \quad (10)$$

There are cases when $v[\] < 0$ is usefully employed: see discussion around Eqn. 13 below. There is also a situation when $v[\] > 0$ is involved, which is when expected mutual information is used, as in Eqn. 8. More generally, now including c , it is considered as the *expected frequency* $e[\]$.

$$v[d, a, c] \rightarrow e[d, a, c] = n[d]n[a]n[c] N^{-2} \quad (11)$$

N is the total amount of relevant data, here $n[d, a] + n[d', a] + n[d, a'] + n[d', a']$. For example, the most basic disproportionality measure is the association constant as follows.

$$K(d; a; c) = e^{I(d; a; c)} = e^{\zeta(1, n[d, a, c])} - e^{\zeta(1, e[d, a, c])} \quad (12)$$

In the case of using Eqn. 11 it is an *atomic association constant* when the reference case of independence in that a , d , and c are all mutually independent, as opposed to, for example, a being independent of a joint event (d, c) where d and c are interdependent. If c is a “joint event” of several conditioning factors, these will also be treated as independent, e.g. as $n[a]n[d]n[c_1]n[c_2]n[c_3]N^{-4}$. It is also called “atomic” because alternative independence assumptions can be calculated as ratios of at least two K .

For purpose of *signaling* an alert for a suspect ADR by a specified drug, only a and d , i.e., not c , are considered. Using spontaneous reporting systems (like FAERS, EudraVigilance, Yellow Card), researchers count how many times a specified drug is reported with an adverse event, then compare this observed count to the expected count (which assumes that drug and event are independent).

The problem with this kind of approach is that statistical sampling of a typical ADR report database is not like the traditional model of sampling as randomly picking black and white balls from an urn, with replacement, nor like analyzing more familiar patient record databases where each patient has a distinct record that occurs only once. As noted in Introduction Section 1, ADR records are duplicated, leading to poor estimates of disproportionality, but do not often contain identical information (and entries can be conflicting), while some duplicates carry a different identifier, looking like a different event. The obvious strategy is to reconcile duplicates into single records, and this approach must be applied in any event when files and hence their records are to be joined. Such deduplication is applied in this study except when stated otherwise, representing cases in which a speedy alert is useful. The second approach is to add further virtual frequencies in the manner of approaches that have sometimes been used in FAERS, e.g., Multi-item Gamma Poisson Shrinker (MGPS) methods (Section 2.5), but now re-expressing this in zeta function form. However, the primary effect is to reduce the strength of the signal for sparse data and a virtual frequency component β “shrinks” extreme ratios toward 1 (the “no-signal” value).

In expressions such as $\zeta(s, n[d, a] + v) - \zeta(s, n[d, a'] + v)$ or its more traditional analogues such as $\log(n[d, a] + v) - \log(n[d, a'] + v)$, where v is the same or similar number throughout, represents a prior belief that here should be no or little difference in counts. This certainly has a damping effect, but it would appear to be in the opposite direction required to remove the effect of duplicate reports statistically, i.e. where it would seem more reasonable to remove an estimate or expectation of the number of duplications. In the popular log LR form, this strategy would be as follows.

$$LLR = \zeta(1, n[d, a] - \alpha) - \zeta(1, n[d, a'] - \beta) + \zeta(1, n[a'] - \gamma) - \zeta(1, n[a] - \delta) \quad (13)$$

This “negative shrinkage” is thus a case in which negative virtual frequencies are used, reminiscent of the use of a Dirichlet “absolutely prior” density of -1 [28,32]. Above, $\alpha = \alpha' n[d, a]$, $\beta = \beta' n[d, a']$, $\gamma = \gamma' n[a']$, and $\delta = \delta' n[a]$ are the prior expected number of duplications (excluding first occurrence) of

records containing “events” $[d,a]$, $[d,a']$, $[a']$ and $[a]$ respectively, strictly meaning that $n[d,a]-\alpha$ etc. are expected counts based on insight as to the statistical consequences of duplicates. It might be well described as a retrospective curation done on a statistical basis. Since α' , β' , γ' and δ' are constants that are each on the range $0 \dots 1$, only, the limits of summation of the zeta function will remain positive. However, this statistical approach is likely to *overestimate* duplicates in quarters with heavy follow-up activity, and *underestimate* duplicates when the same case is submitted by multiple parties with different type codes. But also, raw duplicate counting heavily misrepresents the true underlying number of cases, and it is generally believed (but hard to quantify more reliably), that only 5-10% of ADRs are reported. If so, there are missing counts that should be *added*. Given the above, corrective numbers tend to cancel each other but still lead to very broad error bars on the measures of interest here. Another approach is simply to retain duplicates and consider that the “event” being considered is the report itself, rather than a guaranteed ADR caused by the drug *per se*. That sometimes means that Eqn. 13 is used with $\alpha = \beta = \gamma = \delta = 0$ or a small empirical value. Because reconciling and/or deduplicating is a relatively slow process, those cases needing speed so that deduplication is not applied essentially adopt the above philosophy that the records themselves are the events. In practice, it means subsequently applying more rigorous approaches such as deduplication but only to the drugs to which attention has been drawn. From that perspective, it is not clear that “unhighlighting” signals that arise in sparse data cases (see below) is the appropriate strategy. Rather, an approach such as the above method based on zeta functions to determine the information content of sparse data, on what is argued here to be a more fundamental basis, is indicated.

2.3 Bate's BCPNN Measure

Most detailed comparisons in the present paper are done with FDA methods later recommended for the FAERS database. However, for reasons of data availability and other considerations discussed in Results later below. Bate's method could be considered as foundational because it came earlier than current FDA-approved methods. Not least, Bate expressed the problems, and the solutions, in information theoretic terms. Recall that mutual information underlies Bate's BCPNN measure, $E(IC)$ expressed in bits. The present paper focuses on that rather than with Bate's original neural-net-like component, because when expressed with the zeta functions it requires no arbitrary parameters except as optional virtual frequencies to express prior belief. $E(IC)$ is based on the classical $K(a;d)=P(d,a)/P(d)P(a)$, but as follows.

$$P(d,a) = (n[d,a] + 1) / (N + \gamma) \quad (14)$$

$$P(d) = (n[d] + 1) / (N + 2) \quad (15)$$

$$P(a) = (n[a] + 1) / (N + 2) \quad (16)$$

Here $\gamma = (N+1)^2 / ((n[d,a'] + 1)(n[d',a] + 1))$. Bate [23] proposed $E(IC)$ as follows.

$$E(IC) = E(\log_2 K(d;a)) \approx E(P(d,a)) / E(P(d)) E(P(a)) \quad (17)$$

Strictly speaking, the expectation of a function of parameters is not in general the function of the expectations of the parameters: $E[(x=1:2; y) | D(x,y)]$ (Eqn. 2) is the true expectation of $\ln(P/(1-P))$. “Zeta function counterpart” of the above are exemplified by Eqn. 12 and particularly by Eqn.13 with $\alpha = \beta = \gamma = \delta < 0$.

2.4. The ROR and PRR Measures

The Reporting Odds Ratio ROR approach in ADR signaling [22] uses essentially the form

$$\text{ROR} = P(d, a)P(d', a') / P(d', a)P(d, a') = n[d, a]n[d', a'] / n[d', a]n[d, a'] \quad (18)$$

The ROR is also related to the Proportional Reporting Ratio PRR which is a Likelihood Ratio (LR) and a relative-risk-like measure. For PRR and its comparison with the MGPS Method described next, see Ref [40,42]. Relative Risk is a common example of an LR. The corresponding formula for a PRR is

$$\text{PRR} = n[d, a] n[a'] / n[d', a]n[a] \quad (19)$$

The ROR is also closely related to the Diagnostic Odds Ratio DOR of Evidence Based Medicine. Related to the DOR [43,44], the log ROR (natural logarithm) can be expressed in TEI terms of “estimates” of the natural logarithm, as follows. For comparison with Bate's E(IC), it is re-expressed in units of $\log_2$ in Results Section 4.1.

$$\text{LROR} = \zeta(1, n[d, a]) + \zeta(1, n[d', a']) - \zeta(1, n[d', a]) - \zeta(1, n[d, a']) \quad (20)$$

In terms of mutual information,

$$\text{LROR}' = I(d; a) + I(d'; a') - I(d; a') - I(d'; a) \quad (21)$$

2.5 Multi-item Gamma Poisson Shrinker (MGPS)

The MGPS (Gamma-Poisson) is an ROR approach that “shrinks” the rate ratio as follows, again with expectations taken as the component parameters.

$$\text{ROR}_{\text{Bayes}} = E(\mu_{da} | n[d, a]) E(\mu_{d'a'} | n[d', a']) / E(\mu_{d'a} | n[d', a]) E(\mu_{da'} | n[d, a']) \quad (22)$$

By each one of the four $E(m | n)$ is meant $E(n + \alpha) / (1 + \beta)$, often with a different choice of α and β for each of the four terms in Eqn. 22. It is said to be a smoothed or shrunk estimate of the reporting ratio. β , sometimes called γ prior, “shrinks” extreme ratios toward 1 (the “no-signal” value), working in the count space (Poisson-Gamma). For brevity here, and for comparison with PRR, see Ref [40] and references therein for further details.

3 Methods

3.1 Data, Notation, and Curation

The present study used the FDA FAERS databases [47,48], particularly (but not solely) the DRUG (drug report) and DEMO (patient case demographic) datasets currently with columns headed in the following raw format The ‘\$’ are separators.

```
primaryid$caseid$drug_seq$role_cod$drugname$prod_ai$val_vbm$route$dose_vbm$cum_dose_
chr$cum_dose_unit$dechal$rechal$lot_num$exp_dt$nda_num$dose_amt$dose_unit$dose_form$
dose_freq
```

A typical line of raw data (a record) under the above headers is as follows.

```
100236273$10023627$1$PS$RISPERDAL$RISPERIDONE$1$Oral$$$$$U$Unknown$$020272$$
$Film-coated tablet$
```

“Duplication” is a major issue to consider in a next step of curation. Several records represent the same patient event and the later ones include more detailed, or corrected, entries. Joining

requires deduplication of the DRUG dataset. Joining DRUG and DEMO datasets by primary identifier 'primaryid' generates the following attribute types, here with more standard comma separators.

primaryid,caseid,role_cod,drugname,prod_ai,val_vbm,route,dose_vbm,cum_dose_chr,cum_dose_unit,dechal,rechal,lot_num,exp_dt,nda_num,dose_amt,dose_unit,dose_form,dose_freq,caseid,caseversion,i_f_code,event_dt,mfr_dt,init_fda_dt,fda_dt,rept_cod,auth_num,mfr_num,mfr_sndr,lit_ref,age,age_cod,age_grp,sex,e_sub,wt,wt_cod,rept_dt,to_mfr,occp_cod,reporter_country,occr_country

Major US institutions follow the FDA rule that prefers simple deduplication with selection of later reports over earlier ones as replacements/updates when reporting to the FDA. That FDA rule is, nonetheless, a minimal deduplication rule which is not free of difficulties, and ideally the duplicate records should be joined, entries reconciled, and redundancies removed in one operation. Different institutions use different approaches that can significantly affect signal detection outcomes and reproducibility.

Duplicates are algorithmically filtered during data preprocessing, but there are occasions when more than one duplicate remains. In such cases, if they are not too numerous (see use of a similarity threshold below) they can be treated statistically as independent observations. However, note also that (a) the TEI framework naturally desensitizes signal noise when the surviving records are sparse, and (b) the value of the TEI odds-based approach as used here is that if one has duplicate entries for the same case but the type codes PS etc. disagree, the information is evidence for and against that tends to cancel in a natural way. That is, the overall weight of evidence is reduced by inconsistencies, as if in a court of law. The following method was used in this study and tends to leave just one of the duplicates. DRUG records with the same joining identifier are compared to the extent to which the entries match, which can be controlled by an adjustable similarity threshold parameter so that just very few, typically 1-3, remain. That match test excludes empty entries which the report did not fill in. Then, for those records that survive that process, the record with the most filled-in entries is selected. After some experimentation, the similarity thresholds for results in the present paper was set at 75%. This value was also compared with the results for 100% essential similarity of entries, and for 75% with the same joining identifier not being an essential requirement, because there are duplicates in which a different identifier can be assigned to the same patient event (which the FDA guideline proposal described above would not “see”). Variations in the approach led to somewhat similar LRs for specified drugs with different deduplication strategies including no deduplication. The deduplication and reconciliation process is often by far the most time-consuming of all the steps in analysis. Hence *signal methods* are explored below that can be applied to the DRUG dataset *without* resolving duplicates, but detailed statistical analyses following that resolved duplicates as above. This makes sense in that the initial alert phase essentially focuses attention on a reduced set of drugs of interest.

The steps in curation made use of knowledge element tags in the Quantum Universal Exchange Language (Q-UEL) notation based on Dirac notation, a standard in physics since the 1930s [34,37]. The usual major purposes of these tags include providing statistical summaries [34] and acting as building blocks in inference networks for inference and prediction [37], but they can also act as steps in curation. Indeed, they can also represent a record of *source data* in tidied form, e.g. as follows.

```
<FAER:=DRUG25Q2 'primaryid':='10023627' | if | 'role_cod':='PS ' 'drugname':=' RISPERDAL'
'prod_ai':=' RISPERIDONE' 'val_vbm':=' 1' 'route':=' Oral' 'nda_num':='020272' 'dose_form':='Film-
coated tablet' FAER>
```

Note the metadata operator ':= ' in the attributes. The above examples have a simple metadata *type:=value* structure but using an Attribute Metadata Language (AML) with multiple uses of ':= ' in an attribute would allow flexible interchange with well-structured Longitudinal Health Records using that AML notation [8]. Q-UEL tags are usually read as in subject-verb-object (SVO) languages such as English. The expressions in the Q-UEL tag are logical expressions and by default the logical operator separating attributes in an expression is logical 'and'.

< subject expression | relationship expression | object expression >

Only by useful coincidence do Q-UEL tags resemble XML tags and they can be used in a comparable way. Not least, Q-UEL tags have algebraic and arithmetic force. In Q-UEL programming mode, the tag is treated as the key of an associative array carrying the values. The values are shown up front on the tag as "tag value attributes", when generated, transmitted, or stored on knowledge databases, but are not always seen: by default, the probabilities *Pfwd* and *Pbwd*, association (disproportionality) constant *assoc*, and odds such as *Ofwd* (Predictive Odds) and *Obwd* (Likelihood Ratio) have default value 1 if not specified. The qualification 'fwd' refers to values of the tag being read as written, and 'bwd' to the way it is read when subjection and object expressions are interchanged. This follows principles of quantum mechanics for Hermitian relationships expressions and the tags are used to construct inference net in which causes can be predicted as well as effects (e.g. Ref [37]).

< 'role cod':='nePS' Pfwd:=0.8016 Ofwd:=1.9918 | if:=(assoc:=31.5886, count:=3, factors:=(2,5)) | 'CHECK':='eqcheck' 'occr country':='IN' 'age YR':='30' 'drugname':='MYFORTIC' Pbwd:=0.003280 Obwd:=76.4976 >

3.2 Methods for FAERS Data from the Perspective of TEI

A reasonable information-theoretic construction in the TEI context is as follows, analogous to products of classical log odds ratios. See Section 1.3 for type codes PS etc. There is a change of notation to take account of the role codes, and drugX replaces d for consistency with that and readability.

$$\begin{aligned}
 S(\text{DrugX}) = & \zeta(1, n[\text{PS}, \text{drugX}]) - \zeta(1, n[\text{notPS}, \text{drugX}]) + \zeta(1, n[\text{notPS}]) - \zeta(1, n[\text{PS}]) \\
 & + \zeta(1, n[\text{SS}, \text{drugX}]) - \zeta(1, n[\text{notSS}, \text{drugX}]) + \zeta(1, n[\text{notSS}]) - \zeta(1, n[\text{SS}]) \\
 & + \zeta(1, n[\text{I}, \text{drugX}]) - \zeta(1, n[\text{notI}, \text{drugX}]) + \zeta(1, n[\text{notI}]) - \zeta(1, n[\text{I}]) \\
 & - \zeta(1, n[\text{C}, \text{drugX}]) + \zeta(1, n[\text{notC}, \text{drugX}]) - \zeta(1, n[\text{notC}]) + \zeta(1, n[\text{C}])
 \end{aligned} \tag{23}$$

There are more factors to consider for stratification/segmentation purposes, than just DrugX alone, and we wish to extract Q-UEL tags to provide more detail information relevant to that, so the module DiracMiner includes a form of data mining that can generate many thousands of such tags and combine them in inference to obtain the four Likelihood Ratios with which the following is interchangeable [45,46].

3.3 Processing to Give Simple Empirical Signal Alert Scores

The above formulae could be used to give a value serving as an Alert Signal (AS) in the fashion of FDA and WHO methods but building on Eqn. 23. There is some loss of quantitative information-theoretic rigor because of counting some contributions twice.

$$T(\text{drugX}) = S(\text{drugX}) + \zeta(1, n[\text{PS}, \text{drugX}]) - \zeta(1, e[\text{PS}, \text{drugX}]) \tag{24}$$

It can also be further extended to a measure that subtracts from all the above the corresponding not-drugX case, in which terms not containing drugX will cancel.

$$U(\text{drugX}) = T(\text{drugX}) - T(\text{notdrugX}) \quad (25)$$

Note that zeta contributions can be expressed in terms of Likelihood Ratios (LRs), or the Log Likelihood Ratio.

$$LLR_{PS} = \zeta(1, n[PS, \text{drugX}]) - \zeta(1, n[\text{notPS}, \text{drugX}]) + \zeta(1, n[\text{notPS}]) - \zeta(1, n[PS]) \quad (26)$$

A formal approach to an Alert Score would be to follow Eqns. 23 and 26 and to define an Alert Score (AS) *based on available LRs* is as follows. It is referred to as Type 1 Alert Score.

$$AS_1 = L_{SPS} \times LR_{SS} \times LR_I / LR_C \quad (27)$$

It assumes that the PS, SS, I and C contributions are independent. In practice data for reported LRs can be zero when expressed to very few decimal places, so the following is a more robust measure appropriate for signaling, called Type 2 Alert Score.

$$AS_2 = (LR_{PS} + 1) \times (LR_{SS} + 1) \times (LR_I + 1) / (LR_C + 1) \quad (28)$$

For a simple presentation on an integer scale 0, 1, 2, 3, 4, 5 it is possible to simply add LR_{PS} , LR_{SS} , and LR_I and subtract LR_C and take the nearest integer. To take account of association (disproportionality) constant K as Eqn. 24 suggests, take $K = 1$ above a threshold of significance that is typically 4 and otherwise taken as 0. This gives a Type 3 Alert Score.

$$AS_3 = LR_{PS} + LR_{SS} + LR_I - LR_C + K \quad (29)$$

Another useful measure relevant to signal detection and used extensively in this paper is the statistic here known as Predictability = $LR/(1+LR)$, or Pred for brevity. It arises since in general $LR = \text{sensitivity} / (1 - \text{specificity})$, and at the special “balance point” at which $a = \text{sensitivity} = \text{specificity}$, then $LR = a/(1-a)$ and hence $a = LR/(1+LR)$. Since $\text{accuracy} = (\text{sensitivity})(\text{prevalence}) + (\text{specificity})(1-\text{prevalence})$, the point at which $\text{sensitivity} = \text{specificity}$ is also the point at which $\text{accuracy} = \text{sensitivity} = \text{specificity}$. See e.g. Refs [49,50]. The notion of such a measure is not entirely original as discussed in Ref [44]; for example it is closely related to Lerner’s Q^* discussed there, although the derivation and use is somewhat different. The reason for its limited popularity is perhaps because in specific cases it becomes conceptually difficult, for example because it appears to imply an implausible equality or seems inappropriate as discussed shortly below. In terms of drugX as primary suspect PS, there is a remarkable dichotomy: it provides proof by example of the above while highlighting the conceptual difficulty. $P(\text{drugX}|PS)$ is analogous to sensitivity when seeking to predict drugX as the cause of the ADR from PS-tagged data. Similarly, $P(\text{not-drugX}|\text{notPS})$ is analogous to specificity when seeking to predict not-drugX from not PS. For example, if $P(\text{drugX}|PS)$ calculated from a source data set reflects the probability of encountering records consistent with it in a test set, the analogy appears reasonable. We may write as follows.

$$LR = P(\text{drugX}|PS) / (1 - P(\text{notdrugX}|\text{notPS})) = P(\text{drugX}|PS) / P(\text{drugX}|\text{notPS}) \quad (30)$$

Since by definition of LR, here as $LR = P(\text{drugX}|PS) / P(\text{drugX}|\text{notPS})$, the above is consistent with the previous general description in terms of a and can be considered as proof of it that $LR = \text{sensitivity} / (1 - \text{specificity})$. One point of confusion is in its seeming requirement as occurring only at a special point $P(\text{drugX}|PS) = P(\text{not-drugX}|\text{notPS})$ which looks like an unreasonably restrictive case. In

sampling to obtain these probabilities, how often would such an exact equivalence occur? See explanation below. Another point of possible confusion is that the context is important, but that reduces to a mathematically consistent explanation. If a vitamin taken at correct doses is studied, one finds that $LR = P(\text{drugX}|\text{PS}) / P(\text{drugX}|\text{notPS}) = 0$. However, this means that the use of the prediction rule that the drug will cause a serious ADR will not show any record in agreement, but it clearly predicts correctly when the new LR is taken as $LR = P(\text{drugX}|\text{not-PS}) / P(\text{drugX}|\text{PS})$, i.e., when the LR is less than 1 the reciprocal of the LR is taken as the new LR. That is, for the LR stated, Predictability is given as $1 / (1 + 1/LR)$.

There are strong arguments in favor of Predictability (Pred) as a measure. First, $LR/(1+LR)$ is an indication of the predictive power of the probabilities in the inference net used to compute the LR, although in analogous AI terms, one would raise the objection that the test set was the same as the training set. That transferability to similar cases is plausible is nonetheless a fundamental assumption in statistics, in effect “the point of it”, and the notion of “stationary source” is a fundamental one in information theory. More importantly, it does not prohibit splitting the data into a “training” and test set in some way, and comparing LR or Predictability for the test set. Second, this balance point is often considered the optimal and fairest point on a ROC curve [49,50] and arguably the most important metric derivable from a ROC curve [49,50]. Neither sensitivity nor specificity is favored over the other, and notably it avoids a falsely large accuracy that can be obtained by always predicting yes for a strongly prevalent case (which is why reporting accuracy alone is insufficient). In terms of the present paper, it would be equivalent to balancing a decision threshold to optimize the balance which has the same effect of the case $P(\text{drugX}|\text{PS}) = P(\text{not-drug-X}|\text{not-PS})$ but does not require that such an equivalence is the case.

3.4 ROR* Methods

Originally DOR* methods were developed to predict best options for combining therapies to treat a disease [43,44]. DOR* is the diagnostic odds ratio DOR for a specified disease conditional on a favorable outcome divided by the corresponding DOR conditional on an unfavorable outcome, expressed in TEI form, i.e., incompletely summated zeta functions. For brevity here, reference should be made to the DOR* studies [43,44], but note that, since the unfavorable outcome is of interest as a possible ADR, a corresponding Reporting Odds Ratio is $ROR^* = 1/DOR^*$. An advantage of using the above ROR* is that the possible occurrence of an ADR is not always explicit in available patient data and must be deduced from one or more clinical laboratory results etc. and drug prescriptions, with timestamps [43,44]. Consequently, ADR analytics can be performed on collections of patient medical records without joining ADR report data, which will be described elsewhere. The focus is on the Predictability score based on LR in the present paper.

4 Results

4.1 Step 1 in Developing a New Model. Basic Comparison of Current Approaches and Their TEI Counterparts at the Level of Counts of Observations

As discussed at various points below and especially Discussion and Conclusions Section 5.2, it is not simple to reproduce other methods as they are applied in practice. One may nonetheless see distinguishing features. This preliminary part of the study highlights the differences between

information values (as measured in bits) produced by the above methods by *emulating counts* that keep features detected by the different methods constant, *including total amounts of data* $N=1000$ and $N=100$ and $N=10$, in each of the three cases while association (disproportionality) constant K is increased. K is the basis of Bate's measure, and all results in this Section are converted to log base 2, i.e., units of bits, as in Bate's original work. The difference in behavior between methods is in part due to the emulation algorithm but highlights relevant features. Renormalization to give each value of K allows counts to adopt non-integer values so that the effects of increasing K are plotted as a continuous function, though the points truly corresponding to integer values of counts lie on the curves. The distinction between interpolations and points fitted only shows up at $N \approx 10$. Fig. 1 shows three distinct curves: the methods differ in their favored parameters and by inclusion of different extents of evidence. The MGPS method is one of several based on ROR. That can be shown to be a function of four K s that simplifies, by cancellations by divisions, to a function of four joint probabilities corresponding to the four pillars of evidence.

Fig. 1 explores the $N=1000$ case. Bate's BCPNN approach is shown as the lowest and dark blue curve, and the Bayes-shrunk MGPS approach is shown in green. The brown and light blue curves are respectively their TEI counterparts using partially summated zeta functions. The dotted green corresponds to the MGPS method with an increase of parameters as discussed below. Although all the curves are expressed as log base 2 and hence units of bits here, the brown and light blue curves based on TEI are re-expressed in bits. Natural units or nats are normally used in TEI, since for one observation and no prior information or belief the partially summated zeta function returns a value of 1 nat. The orange-brown curve relates to association as Fano mutual information expressed by Eqn. 8 with a particular choice of virtual frequency $v(\)$ to match adjustable parameters in Bate's method. It corresponds to the expected frequency of observations as discussed immediately following Eqn. 8. The TEI methods used for the light blue and brown curves are, technically speaking, distinct further methods, but are here to show that other approaches can usually be represented in TEI form.

The MGPS approach is that its parameters α and β are adjusted to fit the data: α is particularly important, selected on the Poisson distribution basis. Increasing α and β lifts the curve but reduces its slope. Setting all $\alpha = \beta = 100$ gives 1 bit at $K=1$ and 3 bits at $K=2$, crossing the curve with a and $b = 1$ at *circa* $K=2$ and *circa* 1.5 bits. Adjusting a and b individually for each $E(m | n)$ in Eqn. 22 can clearly exert considerable control to have any required effect. However, for typical higher values in the lower range normally used for each α and β , say 2-5 for α and 1-2 for β , the FDA's MGPS Bayes shrunk ROR curve remains not significantly different from the light blue curve in Fig. 1, but becomes visible behind the light blue curve as slightly lifted and of slightly reduced slope, and still crossing at *circa* $K=2$. The much bigger problem shown by the dashed green line in Fig. 1 is that occasional uses of high values of α and β at *circa* 100 amplify the above "swivel", giving reasonable strong signals for very sparse data, and deemphasizing strong signals. Remarkably, Bate's measure performs similarly to use of zeta functions as shown in Fig. 1, but it is kept in mind that, with the amount of data kept constant at $N = 1000$, the troublesome behavior of the Bate method and the MGPS approach at very low data levels is not illustrated. This issue is addressed shortly below. The consequence of using g makes Bate's measure much less sensitive at least for some reasonable forms of data, with any classical association log constant well above 4 giving essentially the same result as for 4. Evidently, for large N , the log_2 DOR reproducing LDOR' in Eqn. 18, the top, represented by the light blue log curve yield very similar results. Because it sits over the green curve for the MGPS method in Fig.

Fig. 1. Measures of Information in Adverse Drug Reaction Signals. Total data N=1000.

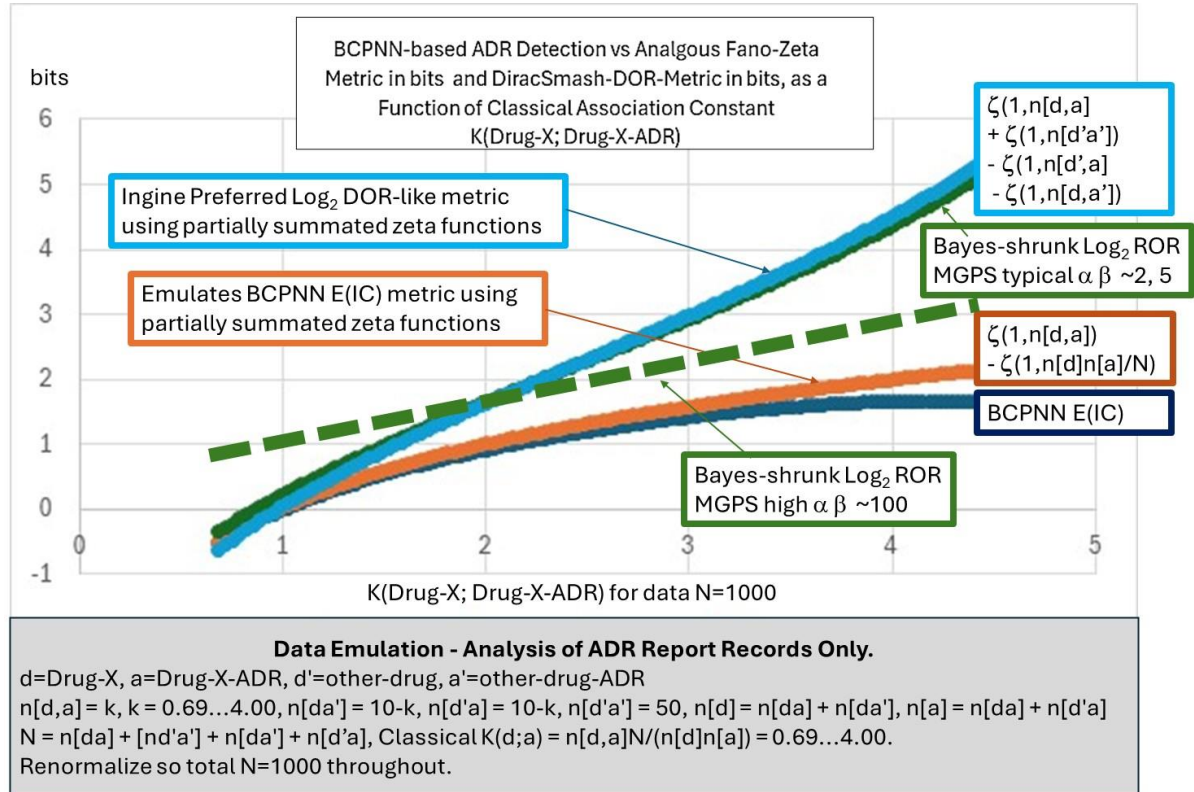


Fig. 1. Widely used ADR statistics (metrics) such as association (disproportionality) constants and odds ratios differ in the details that primarily involve values added to the number of observations and typically involve an empirical aspect to “unhighlight the signals” for very sparse ADR events. They still differ for plentiful data.

1, the blue curve also represents the log base 2 version of the ROR measure and the MGPS approach for very low a and b , which it effectively masks. Specifically, it sits over the log base 2 of the FDA's MGPS Bayes shrunk ROR for lower values of a and $b \approx 1$ (Eqn. 25 and Eqn. 26). However, recall that this is for $N = 1000$ which masks behavior at sparser levels of data. Caution is required when saying that one method carries more information than another, since a value may be lower because contrary evidence, which is more information, is being subtracted. Nonetheless, both the expected information measure and Bayes Shrunk ROR clearly do exceed the other methods in terms of information content and hence, in principle, in terms of predictive power. Converted to LR's for typical association constants found in ADR data, both correspond to Predictability (as discussed in Section 3.3 in relation to Eqn. 30) well in excess of 95%. However, detailed discussion is far from present scope because these are artificial constructs based on preset synthetic counts, very far from "real world data". Moreover, the other methods are intended to be conservative in their strength of signaling.

See Fig. 2, where the total amount of data in the above emulation is reset from $N=1000$ to $N=100$, and then at $N=10$ (when current methods would usually be abandoned or considered unreliable). The situation becomes complicated. In general, when *all* counts $n[d,a]$, $n[d',a']$, $n[d',a]$ and $n[d,a']$ are set equal at 0, 1, 2, 3,... then only the (partially summated) zeta-based measure consistently returns zero, while BCPNN is consistently negative. The log of the basic ROR =

Fig. 2. Different Metrics as Adverse Drug Reaction Signals. Total data N=100 and N=10.

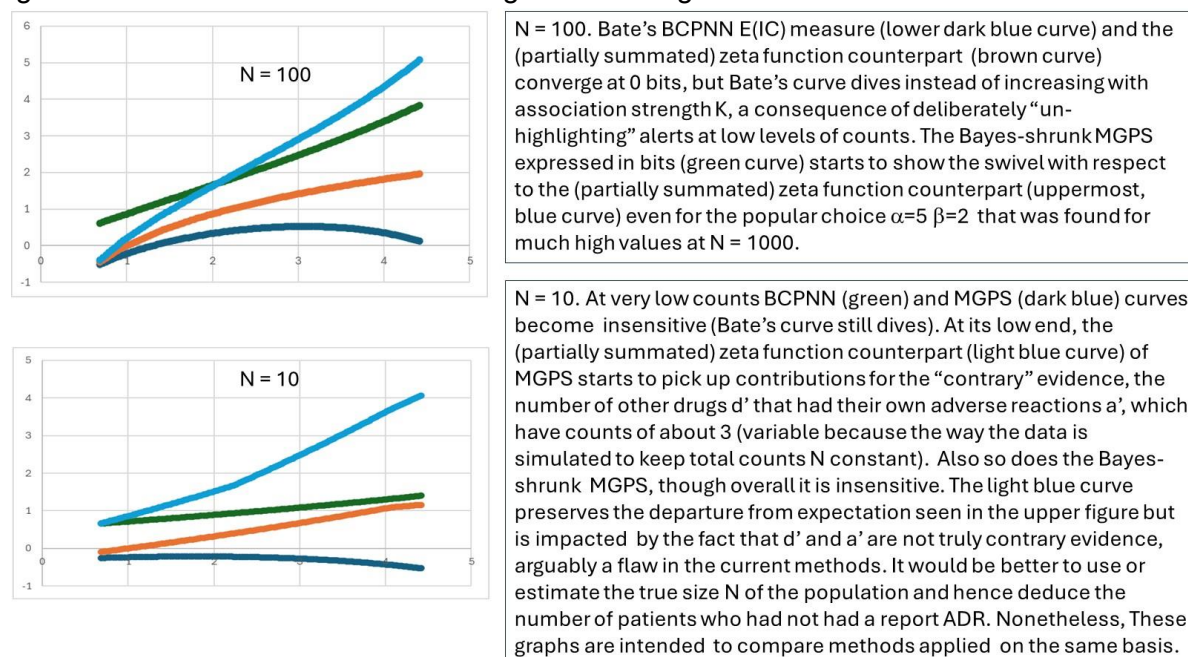


Fig. 2. Widely used ADR statistics (metrics) such as association (disproportionality) constants and odds ratios differ in the details that primarily involve values added to the number of observations and typically involve an empirical aspect to "unhighlight the signals" for very sparse ADR events, at which point the statistics diverge considerably and are usually considered unreliable, hence the "unhighlighting".

$n[d,a]n[d',a'] / n[d',a]n[d,a']$ again behaves similarly to the zeta function approach, notably also returning zero when data is set at equal counts. The value can swing from zero to infinity just from a single report. In general, for sparse data, regulators apply continuity corrections via virtual frequencies, e.g., adding 0.5 to all counts, to avoid infinities. The correction by the Bayesian shrinkage methods of interest here pulls measures toward unity, or zero when expressed on a log basis.

4.2 Tags Generated

See Methods Section 3.1. Tags were generated by mining DiracSmash [45,46] from FAERS DRUG datasets or joined DRUG and DEMO datasets, and from DRUG25Q2 or joined DRUG25Q2 and DEMO25Q2 datasets, including in the preliminary studies. Each tag alone can be considered as a prediction method and contains valuable information, but in much of the present study, as in previous analyses of patient health records [44,45], thousands or hundreds of thousands of these can be assembled by DiracSmash to make an odds prediction network. This allows predictive measures to be used such as Predictive Odds (PO), Likelihood Ratios (LR), and particularly the re-representation of LR as Predictability, i.e. $\text{Pred} = \text{LR} / (\text{LR} + 1)$. For example,

```
< 'role cod':='eqPS' Pofd:=0.9977 Ofwd:=340.8429 | if:=(assoc:=5.2905, count:=1725) |
'route':='Subcutaneous use' 'drugname':='TEDUGLUTIDE' 'drug seq':='few' Pbwd:=0.008724
Obwd:=1466.5683 >
```

Zeta functions were used, but at these levels of data, they would agree with classical ratios of counts to two decimal places. Note that some attribute values can be automatically binned into categories such as 'few' for 'drug sec' (versus 'some', 'many'). The tags used in the rest of this Section and following Sections have the optional "richer" form that make the unsupervised and discovered attributes clearer: the attributes that follow the operator '**with**' which replaces one explicit or implicit '**and**' (logical AND is the implied default). Those attributes following '**with**' are highlighted as found by datamining, and those before it are highlighted as the query. *Efwd* and *Ebwd* are the odds forward and backward for the query alone and have various uses [45,46]. The following examples also illustrate tags resulting from joined DRUG25Q2 and DEMO25Q2 datasets. It also illustrates a negative form to the target (here 'role cod':='nePS').

```
<'role cod':='nePS' Pfwd:=1.0000 Ofwd:=20.0732 Efwd:=0.0222 | if:=(assoc:=45.9478, count:=19,
factors:=(2,5)) | 'caseid':='ge0' with 'reporter country':='CA' 'age':='44' 'sex':='eqM' Pbw:=0.013609
Obwd:=902.2454 Ebwd:=1.0000 >
```

4.3 Combining Tags from Different Sources

Tags containing more attributes can also be formed by tag fusion. For example, a single entry was first extracted from the DRUG25Q2 drug dataset and was joined with the tag from the DEMO25Q2 demographic dataset by a common primary identify (*primaryid*).

```
<'role cod':='eqPS' Pfwd:=1.0000 Ofwd:=2.0368 Efwd:=0.3062 | if:=(assoc:=4.2658, count:=1, factors:=(2,5))
| 'drug seq':='ge0' with 'exp dt':='03513583' 'dose vbm':='300 milligram' join 'primaryid':='160922474'
'caseid':='16092247' 'caseversion':='4' 'i f code':='F' 'mfr dt':='20250407' 'init fda dt':='20190319' 'fda
dt':='20250410' 'rept cod':='EXP' 'mfr snr':='US-SHIRE-US201908573' 'lit ref':='TAKEDA' 'sex':='F' 'e
sub':='Y' 'wt':='59' 'wt cod':='KG' 'rept dt':='20250410' 'reporter country':='US' 'occr country':='US'
Pbw:=4.75041e-006 Obwd:=6.6519 Ebwd:=1.0000 >
```

An example of a joined tag with external information and used in the present project in some investigative studies, is as follows.

```
<Q-UEL-ADR-SIGNAL:=(module:=DIRACSMASH329, comment:=disclaimer:= 'Adverse Drug Reaction
module under development. Accuracy cannot be guaranteed, and no responsibility is accepted for any errors
or consequences arising from their use at this time. ') 'role code':='Primary Suspect' Pfw(P(A|B)):=0.33
Ofwd(PredictiveOdds):=0.49 'Ofwd up to WITH':=0.15 | if:=disproportionality:=1.75 | 'route':='Intravenous'
with 'REFUTE drug as Primary Suspect':=( 'dechallenge':='Y', 'rechallenge':='Y'):=comment:= 'symptoms
abated on discontinuing, reappeared on recontinuing' 'drugname':='CARBOPLATIN' formula:=SMILES:=
'C1CC2(C1)C(=O)O[Pt-2]([NH3+])([NH3+])OC2=O' disease:=cancer Pbw(P(B|A)):=0.000177
Obwd(LikelihoodRatio):=2.09 'Obwd up to WITH':=0.64 'runtime started='Wed Oct 1 12:25:33 2025'
runtime(secs):=67 Q-UEL-ADR-SIGNAL>
```

As discussed in Section 4.3, there is interest in joining ADR information with more detailed patient record information.

```
<'Chronic pulmonary disease':='eqY' Pfw:=1.0000 Ofwd:=2.0368 Efwd:=0.7583 | if:=(assoc:=13.6443,
count:=1, factors:=(5,8)) | 'PO2First OXYGEN PRESSURE mmHg':='eqlow' 'PCO2First CARBON DIOXIDE
PARTIAL PRESSURE mmHg':='high' 'PO2Last OXYGEN PRESSURE mmHg':='low' 'PCO2Last CARBON
DIOXIDE PARTIAL PRESSURE mmHg':='high' with 'Rx:Furosemide':='3' 'MGCnt MAGNESIUM
mg/dL':='high' 'Obesity':='Y' Pbw:=0.00004169 Obwd:=25.7542 Ebwd:=9.5877 >
```

4.4 Even One Observed Joint Event for Adverse Drug Event Carries Information

Note *count:=1* in the above two examples. The ability to handle sparse information at a level relevant to ADR signaling is emphasized by the ability to handle, and have tags for, very few

observations. Even if only one observation was captured in a sample, it is possible for tags to show significant associations.

```
< 'role cod':='eqPS' Pfwd:=1.0000 Ofwd:=2.0368 Efwd:=0.2338 | if:=(assoc:=5.3028, count:=1, factors:=(2,5))
| 'Check':='eqcheck' with 'dechal':='D' 'route':='Intramuscular' 'drugname':='TUBERSOL' Pbwd:=1.0300e-005
Obwd:=8.7640 Ebwd:=1.0060 >
```

It has a significant *association* of 5.3 and an LR of 8.76. This is because it was only expected to be seen $1 / 5.3 = 0.19$ times, i.e., one expects to have needed $1 / 0.19 = 5.3$ times the data (circa 2m) to see it once in circa 10.6 million DRUG records. The probability forward for the order as read is 1, i.e., $Pfwd:=1.0000$, because of the use of the partially summated zeta functions. Although not done in this study, similar considerations can be taken into when tags are missing, e.g. $count:=0$, possible at least for describing relationships with fewer attributes as could be done in CliniMiner [33].

4.5 Step 2 in Developing a New Model. Using ADR Type Code Assignments from Dechallenge-Rechallenge

In contrast to step 1 of model development (Section 4.1) where implied data was artificial, Step 2 addresses real world data. However, PS, SS, I, and C are in some sense still distant from reality. They are subjective assignments. In contrast, dechallenge /rechallenge has the status of an experiment which can be used to test the extent of success of different methods in terms of true and false positives and true and false negatives and measures based on them. See Fig. 3. PS, SS, I, and C are individually predicted with Predictability $Pred = LR / (1 + LR)$ given the conditions that *both* the dechallenge and rechallenge were done and *both* with reported value 'Y' meaning that they were indicative of the specified drug being causal of the symptoms reported as an ADR. Note that SS and I codes are less predictable: they converge less well and closer to $Pred = 0.5$ indicating randomness. The very high Predictability of type code C (here plotted as its complement not-C) does have considerable predictive power and utility but is somewhat artificial because it appears that the reporter usually considered that the dechallenge/rechallenge verifies “experimentally” that the drug in question is likely to be implicated or not implicated in the apparent ADR. PS would seem less determined by this, having a $Pred$ of 0.83-0.85 (say 83-85%) after the first 2 million ADR records.

The finding is that it is reasonable to use PS later below and in future studies, unless otherwise stated, as the primary piece of information, present in the raw data, which serves as the target of interest to predict. There is far more PS data than for dechallenge/rechallenge, and it is a realistic choice of target if it is kept in mind that what it predicts is that the drug will be reported as a primary suspect, with all other influences, e.g. “ecosystem” influences, in mind. When Predictability ($Pred$) is calculated from what is analogous to a training test set, this is analogous to assessing the amount of information, in the sense of predictive capability, that is apparently in that training set. This has been verified by splitting source data of 12 million records into “training” and test sets, providing records are randomly assigned to each set.

4.6 Step 3 in Developing a New Model. Relating Concern Class with Combinations of Impact of PS, SS, I, and C

Although subjective, PS, SS, I and C evidently carry information that is not entirely redundant, suggesting that all should be used in some way. This still relates to an initial focus on signaling as an “early warning” predictive model. “Early warning” refers to cases of particular concern regarding a relatively new drug, or cases of an established drug that might interact with a relatively new drug. See Table 1, which in columns 5 and 6 uses the more familiar LR for comparison with numbers more

Fig. 3. Investigation that a Drug is Causal of the Suspected ADR

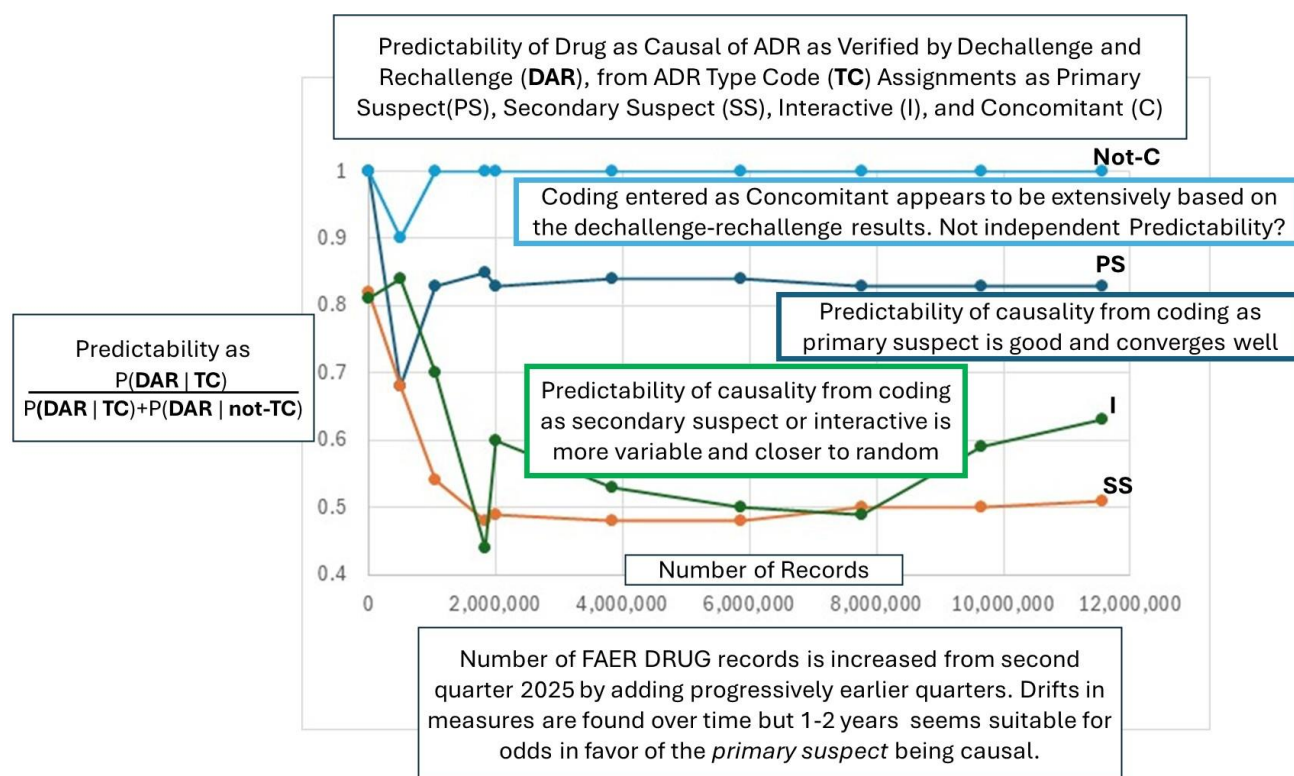


Fig. 3. Predictability ('Pred') is here defined as the point at which accuracy, sensitivity, and specificity have the same value, analogous to a point of balance on a ROC curve. Pred for different descriptions of the involvement of the specified drug in an ADR (the "type code") converges in very different ways as the number of ADR report records in the sample is increased, at least partly because the records appear progressively in time and there are changes in the pharmaceutical ecosystem that affects the number and type of reports.

commonly reported by other workers, and includes data for specific drugs generated in Section 4.7. It is important to note that the drugs mentioned have descriptions which appeared in various sources in 2025 and do not necessarily reflect the correct view today, which is of course also a point on which the approach considered here could be criticized. A reader having any concerns should verify the actual and current status of label warnings, withdrawals, etc. The measures considered as predictive scores are developed from the publicly available FAERS database. However, it is a "real world" approach to prediction outcome, escaping from the limitation that one is using somewhat subjective assignments PS, SS, I and C, to make such predictions. It would be particularly problematic for predictions based on some combination of PS, SS, I and C, since no one of them alone would be a suitable target on which to test predictive capability, and predicting some score from a number of entries would properly require the same combination formula certainly seems to have a cyclic aspect.

Dataset DRUG25Q2 was used here, except where stated otherwise for comparisons (Column 5). Column 1 gives the drug name, written in bold font when it also had a significant association constant greater than 7 for the association of the specified drug with an ADR here determined as assignment of a type code PS by the reporter. The asterisk (*) in Table 1 marks an interesting case. Amoxicillin appears to have had a spate of interaction reports in 2025 that are suspected to be

primarily due to pre-sensitization by β -lactam antibiotics. Column 2 gives some indication of the type of drug, supporting column 1 because one may expect vitamins, creams, and very familiar longstanding drugs for common problems to be relatively safe. Later analyses with more records and more recent records, however, showed an increased tendency to include reactions apparently from overdosing, especially for vitamins A and D. Column 3 gives some description of the state of the drug in the eyes of organizations such as the FDA. Column 4 “Concern Class” assigns a code of 0, 1, 2, 3, or 4, based on column 3 which, as discussed later below, involves some measure of subjectivity and variation of official descriptions with times and source, at least for columns 3 and 4. Column 4 is an indication of the “Concern Class”. This is not the same thing as propensity to ADR, though we may expect correlation with what is indicated by the levels of warnings at the subsequent “label” level and is, of itself, of interest. The Concern Class is assigned as follows.

0 - vitamins, nutrients, creams, or medicine considered very safe

1 - wide use with monitoring or basic cautions

2 - less routine label warnings, typically antibiotics with cross-interactions, or commonly prescribed drugs often quoted as interacting

3 - stiffer warnings corresponding to stratification/segmentation, or previous withdrawal

4 - FDA black box label, a term used by ADR researchers but officially called a “boxed warning”.

Columns 4 and 5 are here regarded as a better indicators of the “real experimental outcome” than use of input states PS (primary suspect), SS (secondary suspect), I (interactive), and C (concomitant) or even the seemingly persuasive use of Dechallenge/Rechallenge in indicating a causal relation between drug and event. This is in view of a surprising weakness of the dechallenge/rechallenge information from Fig. 3 that also shows up in the above column 5. There is rather erratic behavior of the LR based on it. If the LR exceeds 10, it is subsequently set as 10 in further studies below, but its original value is presented in the Table. Nonetheless, Concern Classes 1, 2, and 3 and especially classes 2 and 3 also are somewhat subject to subjectivity in the distinctions made between them, and that even the descriptions that the distinctions are based on can change with time and even vary with information source.

As mentioned above, columns 5 and 6 represent more familiar Likelihood Ratios (LR), here for the Primary Suspect (PS), but with column 6 only for the last quarter, as a very early warning measure. Column 5 relates to runs over the *circa* 12 million records in which a further query was added that the dechallenge or rechallenge or both on records had to have the value ‘Y’, meaning that the named drug should be, or is confirmed to be, a primary suspect PS. The motivation for this was to have a reliable comparison with early warning alerts by having more evidence for the PS state. This is the reason ALECENSA is used as an example above with an LR = 13.83 which had Dechallenge/Rechallenge neglected, so differing from LR = 9.46 in the Table. The corresponding Predictabilities $\text{Pred} = 0.93$ and $\text{Pred} = 0.90$ again emphasize the similarity in terms of significance.

Columns 7 and 8 represent important composite scores of which each combines information from other more elementary scores. Column 7 is the Alert Score Type 3 developed in methods and described by Eqn. 29, and is the earliest approach in this study, and the originally preferred measure for its simplicity. Column 8 is the somewhat more rigorous Alert Score Type 2 which was originally considered only as a intermediate step on the path to Type 3, though in studies below it shows useful stability, i.e., lacks excessive fluctuations. Note that the Alert Scores are not calibrated (parameterized) to reproduce some indication of ADR risk indicated by Column 3 (Concern Level). Rather, they are derived from essentially *ab initio* principles based on LRs (and disproportionality or

Table 1. Early Warning Alerts Compared with Concern Class and Primary Suspect Drug Supported by Dechallenge/Rechallenge Information.

1. Drug. Entries in bold indicate ADR-drug association constant above 7.	2. Description of specific drug.	3. Concern	4. Concern Class	5. Likelihood Ratio for Primary Suspect with dechallenge- rechallenge, Last ~12 million records	6. Early Warning Likelihood Ratio for Primary Suspect Last Quarter	7. Early Warning Alert Score Type 3 AS3 (Original Preferred Measure)	8. Early Warning Alert Score Type 2 AS2 (Current Preferred Measure)
VITAMIN D	vitamin	Wide use	0	0.00	0.0	0	0.38
FOLIC ACID	vitamin	Wide use	0	0.01	0.0	0	0.45
ASPIRIN	NSAID	Wide use	0	0.05	0.0	0	0.92
LEVO-CETIRIZINE	antihistamine	Wide use, monitoring	1	0.19	0.3	0	1.13
DELGOCITINIB	eczema cream	Cautions	1	0.21	0.1	0	0.36
LEVO- THYROXINE	hypothyroid drug	Considered safe classic	1	0.27	0.1	0	0.50
AMLODIPINE	Ca++ channel blocker	Wide use, monitoring	1	1.17	0.4	0	1.35
LISINOPRIL	ACE inhibitor	Wide use, monitoring	1	1.51	0.2	0	0.47
ATORVASTATIN	statin	Wide use, monitoring	1	1.92	0.6	0	1.29
HYDRO-CHLORO- THIAZIDE	thiazide type	Wide use, monitoring	1	0.24	0.2	1	1.74
METFORMIN	biguanide	Wide use, monitoring, commonly “blamed” as other drug present	2	2.64	0.7	0	1.70
CIPROFLOXACIN	antibiotic	Wide use, recognized risk	2	4.10	1.2	1	2.20
AMOXICILLIN*	antibiotic	Caution - sensitized by related antibiotics	2	2.51	0.8	3	5.10
NEMOLIZUMAB	monoclonal Ab	Label warnings	2	2.19	1.9	3	6.96
OFEV	kinase inhibitor	Label warnings	2	2.50	2.3	4	4.62
TEPEZZA	monoclonal Ab	Label warnings	2	2.72	1.4	4	6.48
SKYCLARYS	oxidative stress manage- ment	Important label warnings and precautions	2	3.41	2.7	5	7.03
ELEVIDYS	gene therapy	Previous suspended shipments	3	41.64	3.5	3	4.15

AZITHROMYCIN	antibiotic	FDA warning	3	3.72	1.0	3	5.91
ORSERDU	Chemo-therapy	Label warnings	3	24.86	3.5	4	4.50
ALECENSA	kinase inhibitor	Label warnings	3	9.46	3.5	4	6.30
COSENTYX	immuno-suppress-ant	Label warnings	3	1.48	1.5	4	7.80
BIMZELX	monoclonal Ab	Label warnings	3	9.58	3.5	5	6.30
DAYBUE	neurologic	Label warnings	3	1.57	2.6	5	7.20
SILIQ	monoclonal Ab	Black Box warning	4	3.72	1.4	2	2.18
OBINUTUZUMAB	monoclonal Ab	Black Box warning	4	1.60	1.9	2	3.79
ISOTRETINOIN	antibiotic	Black Box warning	4	9.32	3.0	3	4.62
TRIKAFTA	cystic fibrosis drug	Black Box warning	4	3.21	2.5	3	5.54
INFLECTRA	monoclonal Ab	Black Box warning	4	0.18	0.3	4	4.55
LEQEMBI	monoclonal Ab	Black Box warning	4	15.86	4.3	4	5.30
TRASTUZUMAB DERUXTECAN	targeting by monoclonal Ab	Black Box warning	4	4.50	2.1	4	6.20
TURALIO	kinase inhibitor	Black Box warning	4	(0.39) see text	1.6	4	6.50
ENHERTU	targeted by monoclonal Ab	Black Box warning	4	2.42	2.1	4	7.13
PONATINIB	kinase inhibitor	Black Box warning	4	2.47	3.8	5	5.76

Table 1. Various ADR statistics such as association (disproportionality) constants and odds ratios and further statistics as empirical functions of one or more of them. A patient may be administered several drugs at the same time and the Primary Suspect PS is the one considered responsible by the reporting person. The fifth column compares the cases when the PS assignment is supported by dechallenge-rechallenge, i.e., the ADR declines when administration of the specified drug is halted and resumes on re-administration. These data are for sampled reports in 2025 and a reader having any concerns should verify the actual and status of label warnings, withdrawals, etc. Black Box warnings are officially called “Boxed Warnings”.

association constant K for Type 2) as described in Methods Section 3. Because PS, SS, and I, have only roughly converged, there is emphasis on the contributions that all make collectively, because early warning alerts for new drugs use sparse data. PS, SS, I and C are all given equal weight except that C is considered negative evidence for the drug specified to be associated with the putative ADR. Initially, the preferred Alert Score was Type 3, a function of the five Likelihood Ratios for PS, SS, I, and C and the association constants as described by Eqn. 29, but AS2 (Alert Score Type 2) as described by Eqn. 28 emerges as effective in studies below. Consequently, column 8 is now the one of maximum interest.

Table 1 indicates that the combined model so far is only of moderate quality. It is important to consider Predictability scores in the light of Section 3.3 in connection with Eqn. 30. Considering Concern Class 2-4 as worthy of generating an alert signal, then the LRs for PS Column 5 indicate a range of Predictability (dependent on drug) of 60-96%, after removing Turalio at LR = 0.39, which is evidently incorrect. For Concern Class 1, Atorvastatin seems unreasonably high at LR = 1.92. While the combination model in Columns 7 and 8 are predicting different things, treating Column 8 as a kind of corrected LR suggests 63-89%. Seeing its basis as Eqn. 28 as a product of three LRs divided by one other suggest that a square root would be a fairer estimate of a representative LR on a geometric mean basis, but that would give 57-74%. A more appropriate measure would be how well it predicts the Concern Class. That evidently involves a threshold, which is a characteristic common to the standard methods. Taking a threshold of 2 for the Type 3 model in Column 7 gives an accuracy of 94%, more precisely meaning that it correctly predicts its assignment as either in Concern Classes 1-2 or 3-5 in 94% of cases. Taking a threshold of 2 for the Type 2 model in Column 8 gives an accuracy of 97%, a marginal advantage given the limited number of drugs studied. Going back to the LR in Column 5 and treating it on the same assignment to Concern Classes with a threshold of 2 gives 88%.

4.7 Step 4 in Developing a New Model. Comparison with Standard Methods

As PS, SS, I and C are subjective but still carry information that is not entirely redundant, so do the standard methods, the subjective contribution in this case being implied in the choice of measures, its parameters, and threshold values at which an alert signal is triggered. This suggests that a combination of methods may be fruitful, as was done with PS, SS, I, and C, and similarly an initial consideration involves comparing the information contributions that they make.

Comparison of the methods proposed here with established methods proves difficult. Ideally, one wishes to compare traditional classification metrics: compare Receiver Operating Characteristic (ROC) curves, Precision-Recall curves, or F1-scores. The TEI framework should be benchmarked directly against established methods like BCPNN or vigiRank on a standardized reference dataset (including such as the FAERS, OMOP, or WHO reference standards) to prove it reduces false positives in sparse windows. ROC curves are particularly popular. The present methods including DiracSmash do generate ROC curves, and related comparison metrics, by methods like those in a study with Oxford University and Oxford Drug Design [50]. Perhaps surprisingly, the limitations of the ROC curve approach are not widely discussed, and they are explored in an overall summary and analysis of the results in Section 4.11. Comparisons with various AI methods have also been compared in-house [44]. The problems in doing that for the present ADR study are also some of the major challenges in pharmacovigilance. Even the same declared method for signaling adverse drug reactions (ADRs) is not applied uniformly in practice, and there is considerable ambiguity in how they are implemented and interpreted. There is substantial variation in preprocessing, statistical methods, thresholds, and interpretation. ADR signaling is not applied in the same way across organizations or studies. A signal is issued when odds or information values exceed a specified threshold, but there is no universally accepted value for that threshold. Small differences in preprocessing can substantially affect results. There is variation regarding duplicate removal, follow-up report handling, drug name normalization, coding to preferred terms, and handling missing data. Such ambiguities make blind studies between different institutions difficult. One cannot fully reproduce and compare system behavior without identical data, identical preprocessing, and identical operational rules. This suggests an advantage in running comparative studies within a single institution. These considerations led to the study in Section 4.8. One

can overcome some difficulties by entering at the same levels with the same counts, representing the same number of observations, and this was already done in Section 4.1 and Figs. 2 and 3. The general results for one of the proposed new approaches were described in Section 4.5 One can also compare methods in terms of ranges that embrace the ambiguities and this is done below in Section 4.8.

As in Figs. 1, 2, and 3, one may make some general observations on reported levels of prediction capability. In Fig. 3, there is approximate convergence of prediction of the not-C state (indicating suspected or potential contribution to an ADR), the PS state (Primary Suspect), the SS state (Secondary Suspect) and I (suspected interaction with another drug) at about 1 million ADR report records. This is a convenient number of records corresponding to a typical and a fairly early alert study. Recall that Predictability (Pred) refers to the point at which accuracy, sensitivity, and specificity are equal, which is approximately the inflection point of an ROC curve. Direct predictions of the four states using DiracSmash (Methods Section 3.1) are not the final proposed combination method, but they are insightful. In particular, attempts to improve PS predictions expressed in terms of the accuracy = sensitivity, = specificity balance point are persuasive. On a percentage basis, for the not-C state, the Predictability exceeds 99%, but this was considered somewhat artificial because it appears that the reporter usually considered that dechallenge/rechallenge verified “experimentally” that the drug in question is likely to be implicated in or not likely to be implicated in the apparent ADR, as a strong PS candidate. Nonetheless, it is supporting evidence from the FAER database. The primary Suspect PS performed reasonably well at about 83% and would appear to be the fairer but more conservative assessment. The SS and I predictions performed with a Predictability of around 50-60%, little better than random, and were not studied further. For the not-C and PS studies, same values of more than 99% (99%+) and of 83% are obtained using the analogues of a “training” set and a “test set” in AI studies, both when (a) records sets are randomly shuffled in order and then split, and when (b) the analogues of the training set is the same as the test set. It remains that dechallenge/rechallenge consequences were reported on all records in the analogues of the training and test set, but it may still be considered that this is a valid selection of “quality data”. When the requirement for dechallenge/rechallenge in the training set analogue in the case of PS prediction, 82% was obtained. Training on not-C as combined PS+SS+I in records not including dechallenge/rechallenge data suggesting cause, and tested on those that do contain such challenge information, gave 92%. This value essentially converged on increasing the records in the test, which was considered consistent with Fig. 3. A widespread review of the literature suggests that PRR and ROR methods show *circa* 60-90% sensitivity and 70-95% specificity, and EBGM *circa* 50-80% and 85-95% respectively. Bate’s IC method as used by WHO gives *circa* 50-80% and 85-98% respectively.

Best efforts were made to reproduce these standard methods in house using FAER data, and this gave Predictability scores for PRR of 70-90%, ROR of 70-90%, EBGM 60-80%, and IC 65-80% accuracy at the balance point where accuracy = sensitivity = specificity. An important observation is that changing the selection of source data caused the above dispersion of results but, with the present methods applied to not-C and PS targets, results as described above were essentially stable.

As shown in Fig. 3, a simple direct application of DiracSmash for the PS state converged at 83-85% accuracy = sensitivity = specificity using dechallenge/rechallenge results as indicative of an ADR. Even though this simple single direct application of DiracSmash for a single PS state is not the final preferred method in the present paper (see Section 4.8), not least because it does not necessarily reflect the different and diverse aims of the standard methods, 83% already represents a significant

improvement over the lower ends of the range for other methods which were 70%, 70%, 60% and 65% respectively, and median values of 80%, 80%, 70%, and 72.5% respectively. Also recall that that there was a stringent criterion for a successful prediction (true positive) by DiracSmash: a successful prediction was only counted when both the dechallenge and rechallenge were done, so representing strong evidence that the specified drug was indeed causal of the symptoms reported as an ADR. A less dispersant range of 83-85% over the same changes in selected data would by itself seem meritorious. However, the above ranges for other (established) methods also resulted from seeking to accommodate a variety of difficulties in comparisons as also enumerated in Discussion and Conclusions Section 5, and this requires further extensive study of which Section 4.8 below is a preliminary step. Briefly, to explain the basis for that step, recall that the difficulties in comparison include differences in purpose of methods, i.e., the difference between the objectives of different authors and agencies. One method will do better at another in one test for which it was designed, and that can be reversed in another test. The above issues all represent a fundamental problem addressed next in Section 4.9.

As far as what may be considered as a novel method proposed in the present paper was concerned, methods based on combining DiracSmash predictions for PS, SS, I, and C states with the state C acting as counterevidence for an ADR, were considered intuitively more appropriate and emerged as most successful. Recall that there were three such models tested, represented by Eqns. 27, 28, and 29. Overall, Type 2 combination scores, based on Eqn. 28 and with example values in Column 8 of Table 1 finally emerged as having at the balance point of accuracy = sensitivity = specificity of 96% as assessed by the greatest predictive capability. This would appear to make it superior to the standard methods compared, but there are several reservations. First, while all methods were rightly tested on the same data, this present method has been refined (e.g. by choice of the combinations models in Eqns. 27-29) specifically on the experimental evidence provided by dechallenge and rechallenge and criteria of withdrawal and black labelling as in Table 1, the drugs listed in Table 1, and details of implementation to which there may be objection on grounds of the somewhat differing aims of the standard methods. However, applying the same procedures 5 times to random selections of 50% of 12 million records partitioned into tests sets gave Pred = 93-97%.

4.8 Step 5 in Developing a New Model. Comparison Based on Range of Scores of Combined Methods

Step 4 ran into the difficulty that while different methods contain subjective but nonredundant information, their purposes differ. The main challenge being addressed here is not to show that a prediction module such as DiracSmash is better than another standard ADR method, and certainly not better than all the standard methods combined, because all the standard methods have different aims and often manage different circumstances: they are, in effect, not exactly predicting the same thing. Rather, the issue is seen as one of exploring whether a particular theory of expected information can be introduced into a mix of major types of records to give a more consistent alert signal and predictive capability which embraces the spirits of the methods. Given all the above issues (expanded upon in detail in Discussion and Conclusions Section 5), the position taken is that, for this kind of broad study, one can only scientifically and fairly consider the *ranges* considered by different methods, which from the outset one might have suspected to be quite broad. One can explore how these ranges vary with concern class. This does not mean that it cannot be done for the concern classes as a whole, but it does give further insight and, *a priori*, it could help identify a concern class that is less useful to take into account, presumably because it is less crisply defined. One might, for example, show the correlation with Concern Class of the average over the four Scores in columns 5-

8 for each of the different drugs (all in Table 1). Here, however, the interest is in the ranges shown by all reasonable measures considered, listed shortly below. Only the envelope is shown because the data comprises over 440 points, albeit positioned only at bars corresponding to each of concern scores 0, 1, 2, 3, and 4. They cover all the drugs in Table 1 each involving 7 different alert numbers (including individual LRs used for a similar alert purpose). The data used are

- (1) the Concern Class 0-4 from column 4 as the "x axis", based on the nature of warning labels, withdrawals etc. in column 3.
- (2) the LR of PS supported by Dechallenge/Rechallenge over *circa* 12 million records in column 5, this measure being essentially a measure of the reliability of the drug being assigned as primary suspect,
- (3) the LR for PS for last quarter (DRUG25Q2) in column 6,
- (4) the Alert Score Type 3 from column 7,
- (5) the Alert score Type 2 from column 8
- (6) results for the FAERS data using Bate's BCPNN
- (7) results for the FAERS data using ROR Bayes-shrunk MGPS methods with parameter settings as discussed below, typically 4 distinct numbers for each measure.

However, the LR as reported for the Dechallenge/Rechallenge case (alert number 2 above) is next clipped at a maximum of 10 (prior to averaging) since it proves surprisingly erratic, especially considering it should be a kind of supporting evidence. This is called the Net Score (NS), kept at a convenient range of 0-10, though for the data studied for this figure, the values mostly only reached 9. It is a complex and ultimately still somewhat empirical score of the kind which the TEI ideally seeks to avoid, so it should be emphasized that it is "a work in progress". So far, direct use of Eqn. 23 shows very similar Predictability Pred for type code PS when confirmed by Dechallenge/Rechallenge but appears slightly more sensitive to data over years 2024-2025 than would be explained by changes in the pharmaceutical market ecosystem.

The Net Score is intuitively reasonable as a compromise. That is, reasonable as consistent with an average of equally weighted scores of several different measures that are individually much less empirical and primarily LRs for different type codes and conditions such as Dechallenge/Rechallenge and different amounts (time periods) of data. The explanation for the height of each range, i.e., the width of the envelope of the band shown for blue bars, is thus that the measure is obtained for each of many drugs. However, the five vertical blue bars also contain most of the drug scores (i.e., the spread over the individual four measures each done for a different drug), each measure behaving rather erratically with respect to each other at a finer level of detail but still showing the same overall trend. Notably, for alert signals, (a) almost all drug scores (LRs etc.) for Concern Class 0-1 have values in the range 0-1, (b) almost all drug scores in Concern Class 2 have values in the range 2-6, and (c) the majority of drug scores in Concern Classes 3 and 4 have values in the range 3-8, the main exceptions turning out to be the more erratic measures conditional on Dechallenge/ Rechallenge. Overall, however, Likelihood Ratios (LRs) for the data used tend to adopt similar ranges of values as the Type 2 and 3 Alert Scores, so that the LR itself can be considered a kind of alert score.

The 5 vertical blue bars at Concern Scores 0-4, represent the spread of different reasonable measures that can be used for alerts, including the use of DiracSmash generating LRs as described above, and this highlights that all the measures used do, despite considerable differences, correspond to the use of Concern Classes, which they predict. Of particular importance here are

the spreads of measures from Bate's BCPNN and ROR Bayes-shrunk MGPS methods with different choices of parameters described in various sources, as noted above and discussed in more detail below, because these are standard methods in the field. The horizontal dashed lines represent commonly used threshold values for signals of different strengths, essentially meaning of different sensitivities and specificities. There is no great sense in which one threshold can be considered more correct than another, but it seems clear that, by lying at the top of the range of major method scores in the blue bar, Bate's "typical strong signal" (red dotted line) for an ADR of great concern in Concern Classes 3 and 4, is reasonable. To compare scores on a more analogous basis, the BCPNN and ROR values were expressed as unlogged ratios (e.g. 2 to the power of Bate bit values) with a range of typical parameter settings. Also, as noted in Figs.1 and 2, "unhighlighting" to avoid signals by these methods when sparse data complicates comparison, so comparison is preferably made for sufficiently larger data, say above 20-50 events. For the data used and ranges of parameters believed to be most typically used in the approved methods, all fall in the envelope shown, but in principle, Bate's BCPNN and ROR methods can greatly increase above the values of range of scores studied in Fig. 4. All the above are plotted as functions of Concern Class 0-4 because this is, as noted above, surprisingly a more stable "experimental result" for validation than Dechallenge-Rechallenge, albeit retrospective. Of course, classifications PS, SS, I, C dechallenge/recallenge, and the concern classes are not perfect and so contribute to dispersion. The ranges represented by both the red and red plus blue vertical bars in Fig. 4 are those for *typical* FAERS DRUG data for each Concern Class. By "typical" is meant that they are measures from the last available quarter as tests of early warning signal scores, but also the LR for code type PS for 6 quarters (which is circa 12 million records) for comparison, and this latter measure contributes significantly to the dispersion.

The Concern Class seems the most plausible "outcome" measure for testing alert signals, but it has a strong subjective component both in the way authorities assign the description and its scope on a warning label, and in interpreting the impact of label warnings. Overall, based on Bate's and ROR thresholds corresponding to score 1 for a "mild signal", 95% of predictions done by the present TEI and the standard methods were correctly assigned to Concern Classes 0-1. Based on Bate's and ROR thresholds corresponding to score 2 for a typical "medium signal", 95% of predictions done by the present TEI and the standard methods were again correctly assigned to Concern Classes. These can be related approximately to Predictivity 95%. Other common choices indicated by the remaining horizontal dashed lines greatly reduced Predictability. All the above measures, and hence Fig. 4, are useful for summarizing the overall impact of ADR reporting, but they do not capture information from important counterfactuals, for example that what is perceived as an ADR is a set of symptoms and/or clinical laboratory results are due to some other cause than the drug specified. Nor do the measures, including the concept classes, measure the benefits of a drug for rarer disease when alternatives are not possible. Such assessments are best done by consideration of Longitudinal Health Records, as discussed briefly below.

4.9 Step 6 in Developing a New Model. Attempted Refinement as Signaling and Analytical Methods

Despite variations and uncertainties, a proposal can be made as to best approach based on the new Signaling method described so far, as Fig. 4 also illustrates. It is Concern Class 3 on the y-axis that emerges as most diverse in terms of types of information used for that assignment, subjectivity in interpreting that information, and in terms of the variation of that information with time and source. If the contribution of this Concern Class is removed, which is in effect the same as bringing it into alignment with Concern Classes 2 and 4 by assuming that the relation should be

Fig 4. Comparison of Various Early Warning Alert Scores with final drug Concern Class assignments as proven almost completely safe (0) compared with FDA Black Box warnings (4) and intermediate indications.

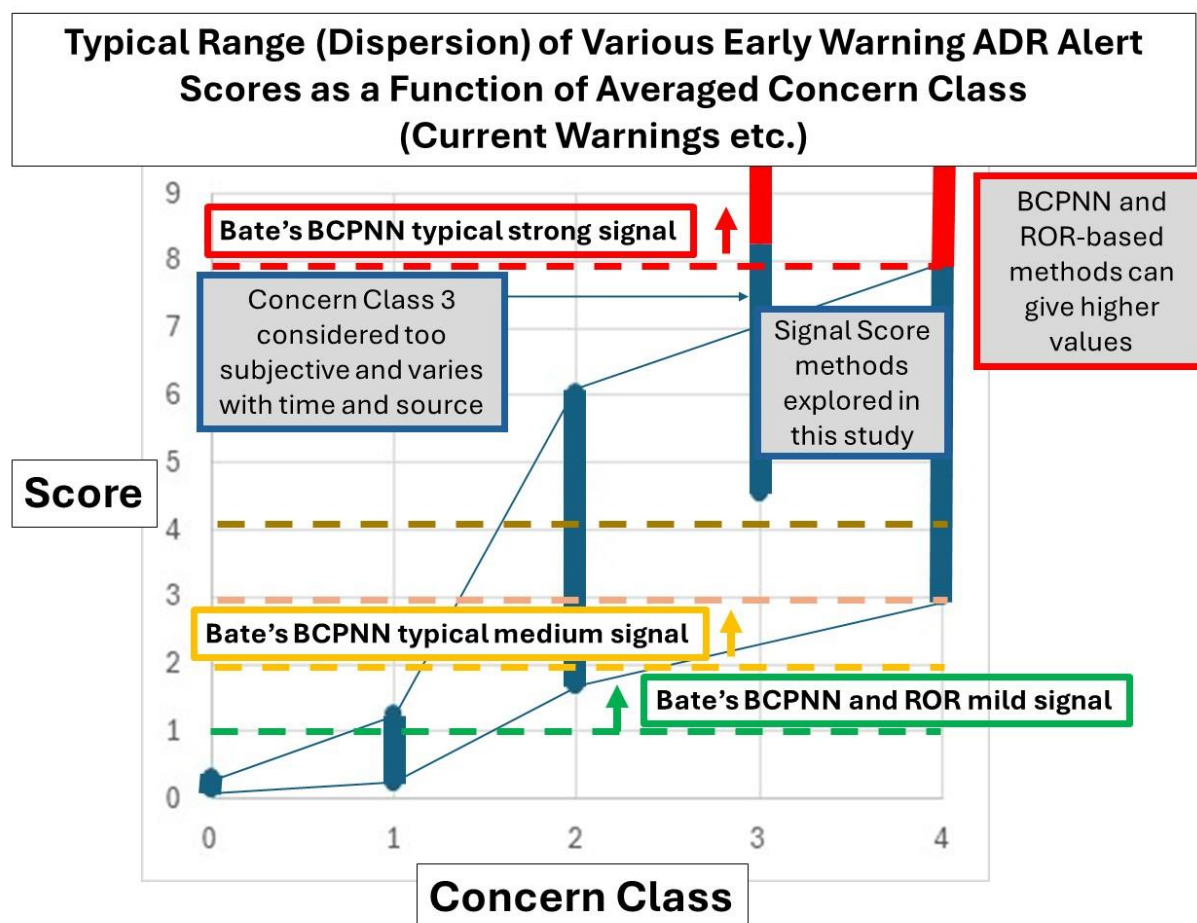


Fig. 4. The various statistics (metrics) used as ADR signals can be compared with the historically observed outcome as a “Concern Class”, i.e., a measure of the risks of the ADR for a specified drug versus the benefits that manifest as contraindications and warnings on labels, “black box” labels (a more extreme case enforced by the US FDA), and withdrawal. These statistics cover a broad band not least because in this case they relate to a large sample of different drugs as in Table 1. The envelope, however, suggests an appropriate ADR alert metric from a combination of the different statistics.

monotonic, an interesting envelope emerges, represented by the thin blue lines. The lower blue line might be taken as a functional form such that above the line a signal should occur, but overall, there is a natural choice of thresholds as follows.

- Alert Threshold 1, at Net Score 1, green dotted line, also the mild straight signal for Bate's BCPNN and ROR methods.
- Alert Threshold 2, at Net Score 2, yellow dotted line at score 2, also the medium straight signal for Bate's BCPNN for typical data as defined above.
- Alert Threshold 3, at Net Score 3, light pink-brown dotted line, also the level above which drug scores fall into well-defined Concern Class 4, i.e., the FDA Black Box label assignment.

- Alert Threshold 4, at Net Score 4, dark brown dotted line, also the level above which some drug scores fall into Concern Class 4 of Black Box assignment or a more poorly defined Concern Class 3 of more severe warning, stratification/segmentation, i.e., the FDA Black Box label assignment.
- Alert Threshold 5, at Net Score 8, red dotted line, also the level above which some of Bate's BCPNN method tends to assign as a strong signal and ROR methods tend to do so, for this kind of data.

Even though the above focuses on the signaling aspect, it remains that the method described here, from the outset, went a little further along the continuum to include analysis, if only better to qualify the signal, i.e., to ensure that the initial signal is potentially real enough to be worthy of attention. This analysis can be further extended using the current proposed approach, as follows (Sections 4.9 and 4.10). The simpler form of the envelope improves prediction of very strong signals as opposed to milder signals but what constitutes a very strong signal is a somewhat artificial, albeit useful, definition. It gives the distinction between strong and milder signals a more rational form. However, it does not affect the performance with Predictability of 95% as an alert signal based on minimum triggering thresholds of 1 and 2 discussed in Section 4.8.

4.10 Predictability of ADRs from Basic Demographics

The above is a combination approach based not only on a variety of new measures explored here but on two major established methods of the FDA and the WHO. There is a further approach to analysis that is peculiar to the authors' general approach and thus only represents the use of the new methods described. An unusual feature is that the Q-UEL language used here provides annotated probabilities, knowledge elements that can be analyzed and reused. Each Q-UEL "tag" is itself a prediction even prior to being assembled in a prediction network, but an exact calculation of probabilities except for dependence on amount and quality of the data, in contrast to any prediction method which is essentially an estimate, primarily due to independence assumptions. Predictabilities from basic demographics come from single Q-UEL tags (Section 4.2). See examples from Figs. 5, 6 and 7. Joined DRUG25Q2 and DEMO25Q2 data sets are used here to illustrate capability from Early Warning studies. The ability to compare predictabilities with those obtained above properly depends on integrating ranges under these curves, and the contributions from each factor should really be combined, i.e., treated as joint events, but it is clear that the probability of an ADR is enhanced around certain values of age and weight. A simple preliminary study predicting LR *based on predicting PS* followed by assignment based on LR into two grouped Concern Classes 0-1 and 2-4 by DiracSmash [45,46] based on sex, age ranges 0-9, 10-19, 20-29 etc. and weights based of 30-39, 40-49, 50-59 etc. kg gave an LR of 32.3 and a Predictability of 97% for the drugs in Table 1, a small but significant enhancement over 95% achieved by the methods above. This needs improving in various ways, including by focusing on dechallenge/rechallenge results as more persuasive evidence.

In Fig. 5, an age around 70 years is a strong influence while younger ages are influenced by other, perhaps as yet unclear, factors. However, it also reflects the distribution of age groups most affected by ADRs. This example includes the effects of many specified drugs combined.

4.11 ROC Curves and Appropriate Measures for Comparing Performance.

Comparative performance of different methods is always of great interest, and the methods used above for assessing predictive quality have been somewhat unusual. First note that the method of assessing predictive performance by means of, e.g., signal alerts, is not in general the same as

Fig. 5. Dispersion and Maximum Predictability by Age and Sex.

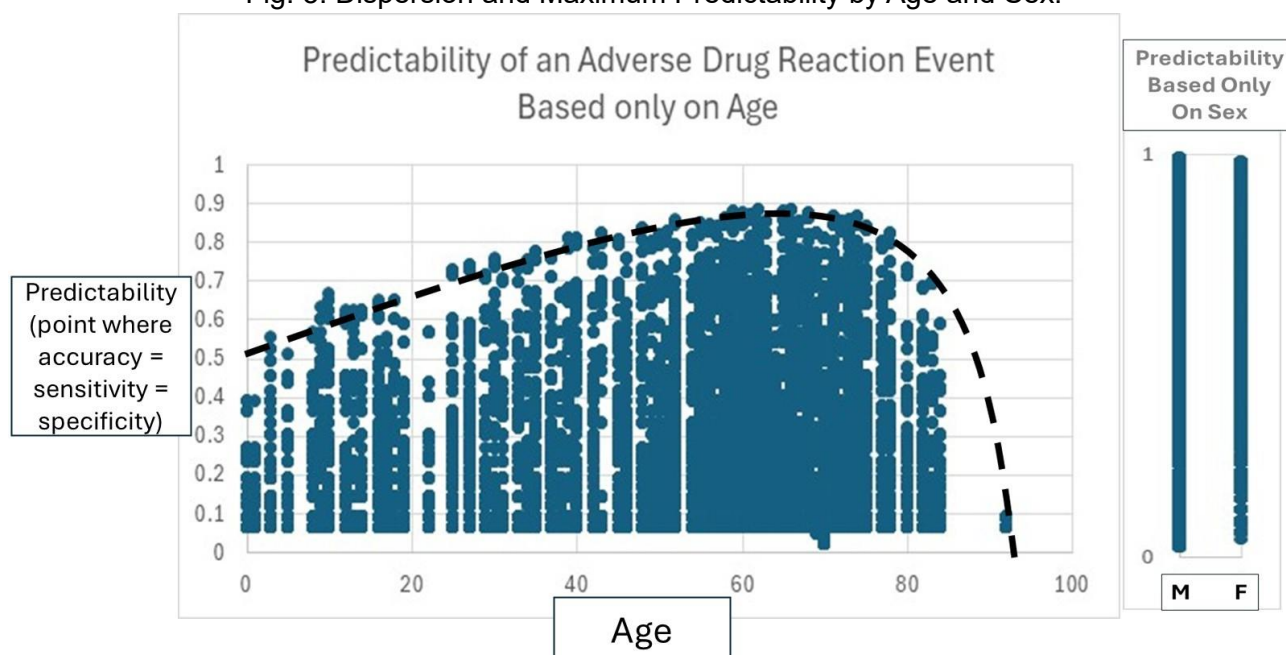


Fig. 5. A distinctive feature of the datamining approach used in the present paper is that the joint events for the dependence of each named drug with conditioning factors such as patient age, sex, and weight and other factors are captured as a large number of reusable “knowledge elements” (“tags” expressed in the Q-UEL language) that can be combined to form inference networks that make predictions. However, they also individually represent predictions with different degrees of accuracy etc. and the varying degrees of that for one or more conditional factors such as age suggest the impact of other influencing factors. Other examples are shown in Figs 6 and 7.

measures used such as Bate’s method based on mutual information, or the ROR, or in the present study a formal TEI approach (Eqn. 23) or simple empirical estimates based on combining PS, SS, I, and C (Eqns. 27-29). However, a performance measure could be used as a signal alert, and Likelihood Ratio LR for PS appeared as one of the alert signal methods considered in Table 1. The LR is a popular measure of predictive performance as well as a statistic and a measure in Evidence Based Medicine, Relative Risk being a specific example of an LR. Though the method frequently used in the present paper is somewhat unusual, it is more a matter of reinterpretation of LR, being derived directly from the LR, as Predictability Pred = $LR / (1 + LR)$ interpreted as the value of accuracy = sensitivity = specificity, i.e., at their balance point (see Section 3.3).

Before proceeding on to Section 4.12 which is a summary of the all the results so far in terms of comparative performance, including those appearing in steps to develop an improved model, it is helpful to present results of studies comparing the performance measures themselves, and comment on how they relate to ROC curves, Precision-Recall curves, and F1-scores which are commonly used by other researchers in pharmacovigilance. Precision-Recall curves are arguably more appropriate to current mainstream Machine and Deep Learning and suffer from some of the same issues as ROC curves discussed below. They are dominated by the trade-off between recall (sensitivity) and precision. Precision is the fraction of predictions of e.g., PS that are correct. It is prevalence-dependent while the Likelihood Ratio LR, and hence Predictability, is deliberately not. The F1 score is related to the ROC curve insomuch as the ROC curve describes the performance of a classifier

Fig. 6. Dispersion and Maximum Predictability by Weight.

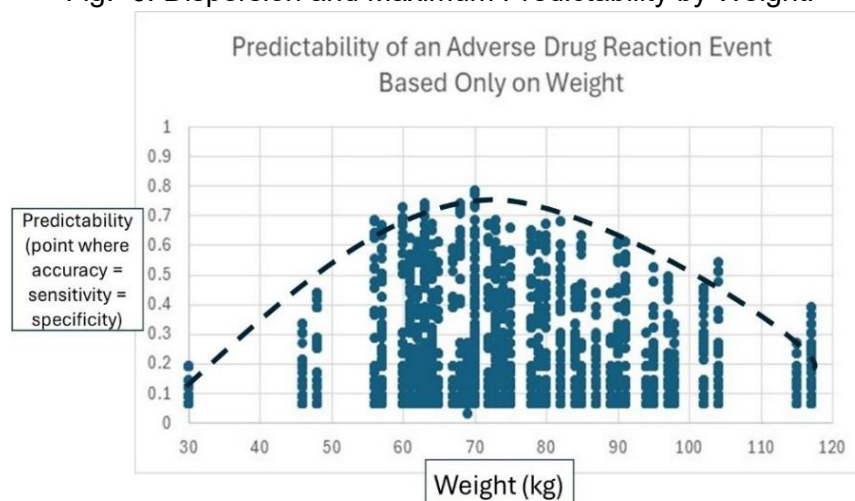


Fig. 6. As for Fig. 5, a distinctive feature of the datamining approach used in the present paper is that the joint events for the dependence of each named drug with conditioning factors such as patient age, sex, and in this case weight can be obtained from Q-UEL tags. While few report records specified weight, it is clear that a patient weight of around 70 kg gives strongest predictions for records in recent time.

Fig. 7 Dispersion and Maximum Predictability by Combinations of Key Demographic Factors.



Fig. 7. As for Figs. 5-6, a distinctive feature of the datamining approach used in the present paper is that the joint events for the dependence of each named drug with conditioning factors such as patient age, sex, and weight and other factors can be analyzed from the Q-UEL tags. This figure gives some examples of combinations of two such conditional factors at a time, and also an example of predictions of the combined non-concomitant type codes, i.e., that the drug is a primary or secondary suspect, or believed involved in interactions with other drugs.

over all possible thresholds and the F1 score is a single-number summary at one specific threshold, although it represents the geometric mean of current AI's precision and recall. The LR is an

exponential of an information measure in natural units that measures how much a finding changes the odds of an outcome, rather than simply whether a prediction is correct. This has several important advantages when comparing prediction methods, especially in settings like FAERS pharmacovigilance. There, outcomes are often rare, data are often imbalanced, and decisions need to depend on the strength of evidence which is the essential purpose of the TEI approach. There is a sense in which the LR and hence Predictability is the more general, and even the effective mother, of the above alternatives. If records are ranked by decreasing LR as e.g., $P(\text{drug} \mid \text{PS}) / P(\text{drug} \mid \text{not-PS})$ then each threshold value for LR discussed shortly below, also defines one ROC point, one Precision–Recall point, and one F1 score. This means one can construct F1 as a function of the likelihood-ratio threshold. Comparing the threshold that maximizes F1 with the thresholds for LR = 1, maximum Youden index (sensitivity + specificity - 1), and sensitivity = specificity provide insight into how different decision criteria behave for a pharmacovigilance model. This is considered important for imbalanced datasets like FAERS, where these thresholds often differ substantially.

ROC curves and related measures [49] are a very popular choice for comparing prediction methods in general including by the present authors and collaborators [50], and still so in ADR analytics. This might be because they are easier to understand than the alternatives, rather than because of the advantages. However, they must be used cautiously, and this is not too often done in literature, so this should be discussed first. Recall that frequently throughout the present paper so far, Predictability derived from the LR has been used to assess information content relevant to the potential predictive performance of a method, as well as being a useful statistic (e.g. Section 4.10). The purpose is also to summarize the essential features of a ROC curve for prediction. See discussion around Eqn. 30 in Methods Section 3.3. Nonetheless, it is preferred that this method is used rather than the ROC curve itself or related measures such as Area Under the Curve, not only for reasons discussed in Section 3.3, but because of the following. The tracing of a ROC curve is based on the explicit or implicit use of a decision threshold varied from sensitivity of 0 to 1 (or 1 - specificity = 0 to 1 - specificity = 1) which maps to, i.e., effectively traces along, the ROC curve. The real prediction should be based on a threshold value, corresponding to a point on the curve, which is part of the model developed in the analogue of the training set, not at the time of it being applied to the test set. Basing success on the ROC curve as a whole or area under the curve implies prior knowledge of the answer (i.e., the positives TP and FP and negatives TN and FN). The best that might be said is qualitative, that a “fat” ROC curve produced for a test set (and training set) increases the *probability* of a predictive model being useful. In contrast, Predictability as the balance point where accuracy = sensitivity = specificity depends only on the Likelihood Ratio LR and the only issue is whether it is being estimated for a dataset that is used to derive the model, i.e., the training set, or tested on a randomly selected subset of records available to serve as a test set.

Being dependent only on the information content for prediction represented by the single quantity LR (the exponential of the relevant information in natural units), the approach tends to give similar measure of Predictability and ROC curves for randomly selected records as “training” and various test sets. The seeming considerable variance in the literature is extensively due to ambiguities in implementation, parameter choice, and thresholds (see Section 4.7). In the studies above, providing that the number of records used was minimally about two million and preferably 12 million, and more precisely if each drug being addressed is represented by at least roughly 400-500 records, the model is transferable from the training set with a Predictability per cent of $\pm 2\%$ standard deviation which seems surprisingly small but is, at least for PS, consistent with Fig. 3. That is, also

providing that training and test dataset records are truly close to being randomly chosen and not too far apart in date so that the pharmaceutical ecosystem is comparable. Predictability is also a fair choice because any one of accuracy, sensitivity or specificity alone can be misleading when what are considered as positive and negative outcomes are far from being in equal prevalence.

Fig. 8 shows as the upper blue curve a typical ROC curve for Predictability of 95%, which was characteristic of model development steps 5 to 7. The large points are the Predictability deduced from the LR and the smaller points correspond to ROC curve studies with the nature of the adjustable threshold discussed below. For $1 - \text{specificity} \geq 0.3$, values for the smaller points are emulated to interpolate between check calculations but since a ROC curve is inevitably monotonic concave in practice, this necessarily represents the proper form. Not all drugs in all situations and modes of analysis achieve this high score. The Predictability and ROC curve performance of a specific drug based on prediction of PS is well illustrated from the example of Ciprofloxacin shown, which shows improved performance for predicting it as Primary Suspect (PS) when the analogue of the training set is increased in size and more so when the outcome of dechallenge and rechallenge is being predicted, or only records carrying dechallenge and rechallenge data are used. However, this increase with more ADR records is primarily a matter of convergence, as Fig. 3 showed. In some ways this variation is rather surprising, as for the red curve for one year quarter there are 428,772 records with PS, 1,400,284, for SS, I, or C, and 474 instances of Ciprofloxacin as PS, seemingly a reasonable amount of data by some ADR analytics standards, at least if the last very few quarters are examined or the drug is rarely used. It may suggest that much variation is due to the changes in the pharmaceutical ecosystem, e.g., possibly due to new drugs being introduced, some with cross-reactive effect. The actual case used for the green curve in Fig. 1. is as described in Section 4.5 and Fig. 3, which relate to part of Step 2. DiracSmash [45,46] can be used to assess other selected points of a record, e.g. as described in Ref [50]. The 12 million records carried 2,507,906 cases of a drug being a Primary Suspect of which 3,155 related to Ciprofloxacin, and further data mining for possible influences generated 82,542 “tags” for estimation of the final LR = 1.5026. An example tag for that case is shown below. Recall how the LR is computed by DiracSmash. Each tag represents a single knowledge element which alone can be considered as an inference network analogous to a Bayes Net, but now with bidirectional conditionality and based on exact counts with no independence assumptions. In this case, the probabilities and odds were estimated with zeta functions (using $s=1$).

```
< 'role cod':='eqPS' P fwd:=0.7380 Ofwd:=2.6402 Efwd:=0.4157 | if:=(assoc:=3.4059, count:=44,
factors:=(2,5)) | 'drugname':='eqCIPROFLOXACIN' with 'dose form':='MG' 'dose amt':='77859' 'rechal':='D'
Pbwd:=1.7973e-005 Obwd:=9.5442 Ebwd:=1.5026 >
```

Many authors assess prediction method using LRs but it can be argued that an LR is not a single kind of measure. LRs can differ in the nature and number of independence assumptions made in their calculation. An assumption of independence is an assumption of random association. All prediction methods use such assumptions when insufficient observations are available for exact calculations, otherwise calculations based on exact counts of joint events would be done. Essentially, use of independence assumptions are their purpose. Conversely, by “exact” here is meant subject to limitations of data, but probabilities such as $P(A,B,C,D)$ are based on counts $n(A,B,C,D)$. An important feature of DiracSmash is that it can control the use of independence assumptions. The whole inference network generated by DiracSmash, which is an odds inference net [45,46] based on

Fig. 8 Relationship between the Predictability Measure Used in the Present Study and ROC Curves.

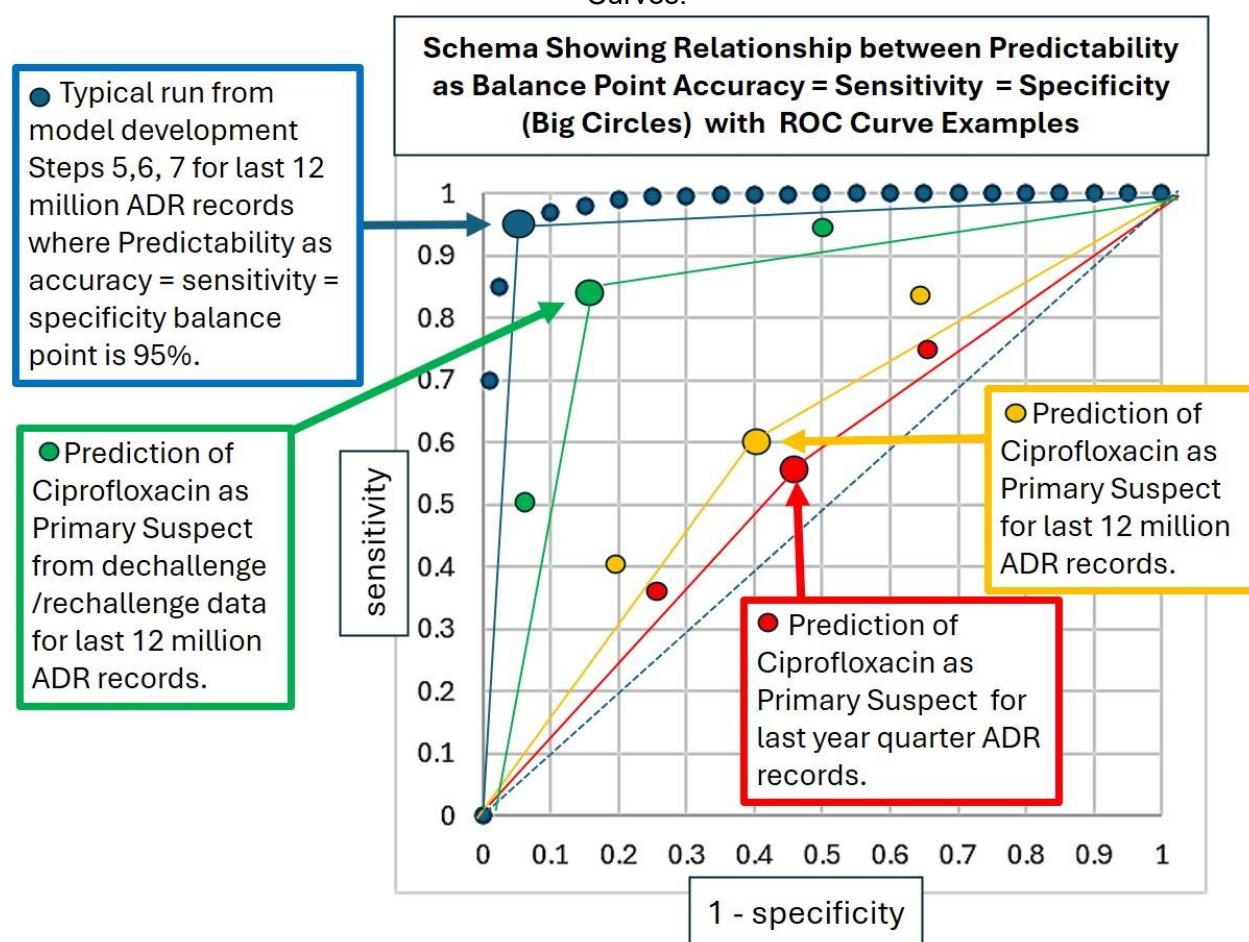


Fig. 8. The relationship between Predictability (Pred) as the important balance point accuracy = sensitivity = specificity and the ROC curve. Predictability lies on the diagonal from sensitivity = 1 to 1-specificity = 1. The Predictability points are sufficient to capture, to good approximation, the essential features of the ROC curves as typically calculated.

combining many tags like that above, necessarily assumes independence between attributes such as 'rechal':='D' when tags do not contain the same attribute. This is the same concept as $P(A,B) \times P(C,D)$ being an estimate of $P(A,B,C,D)$, with (A,B) and (C,D) being treated as mutually independent even if they are not. The overall LR computed from the odds inference net when using such independence assumptions because of limited content in tags was 2.048. However, DiracSmash enables the user to control and minimize use of independence assumptions and so obtain greater accuracy. In this calculation, DiracSmash was set to calculate interdependencies from exact counts for parts of the calculation by the network for which this was possible. Because the amount of data represented by 12 million records is more than sufficient for the simple case of a PS assignment given Ciprofloxacin as the drug, i.e., only two attributes, the LR obtained by the entire adjusted network matched the LR that would be obtained by exact counts using a standard approach. The role of the tags was, in that sense, reduced to giving insight into impacting factors that could be explored. The data levels in this case are also sufficient that the use of zeta functions rather than classical odds made little difference, because classical methods and TEI converge when data is

plentiful. This exact solution based simply on counting multifactor joint events is not always possible and is not a common use of DiracSmash (except at the level of each tag generated) but is so for the green ROC curve of 12 million records. It corresponded to $LR = 1.503$ as stated earlier, slightly lower than the above 2.048 obtained using unnecessary independence assumptions. These LRs represent Predictability scores of 60% and 67% respectively, in accord as a qualitative agreement for a weak positive prediction, though exact calculations, when possible, are preferred for epidemiological and public health resource calculations, hence the above strategy used by DiracSmash.

The adjustable threshold for ROC curve generation arises as follows. For the exact $LR = 1.503$ case, Prior odds $PrO = 0.277$, and Predictive Odds $PO = 0.416$ were obtained. Recall that one way of thinking about the LR is that it is the PO for the specific drug compared with (divided by) what would be expected on a chance basis, which in this case is considered as the Prior Odds. This also represents an adjustable belief or external information, and importantly it can also be considered as the adjustable threshold in a ROC curve. Basing the threshold on Prior odds means that a prediction, e.g., that a drug is a primary suspect PS, is made if $LR > 1$. For tests on many records, if a record matches (contains) the factors, e.g., demographic and clinical factors, in the query for supervised data mining (in this case by DiracSmash [45,46]), and the LR exceeds the threshold, the prediction is taken as positive. It is a true positive TP if the record indicates that it fits the target, e.g., that the drug is a PS, and a false positive FP otherwise. If a record does not match them, the prediction is negative. It is a true negative TN if the record indicates that it is not a PS, and a false negative FN otherwise. DiracSmash does allow that selected conditional factors can be assumed independent, which would suggest the option that only at least one of them is in a record, but there were no such independent factors in the Ciprofloxacin case, and a ROC curve for the case of 95% Predictability is not materially different whether or not independent factors are rated this way. The sensitivity and specificity are calculated from the numbers of TP, FP, TN and FN. The small points indicate these measurements.

The points for Predictability, and for sensitivity = 0 and specificity = 0 tell much about a complete ROC curve given also that it must be convex monotonic for adequate data and any reasonable predictive model. As a first approximation, a triangle defined by these three points is an adequate indication of trends in area under the curve in comparative studies. However, it does not necessarily follow that Predictability is the point of inflection, i.e., the peak when the curve is rotated 45° clockwise. If Predictability and inflection points are similar, so that the curve appears symmetrical about a diagonal from top left to bottom right, it is usually taken as indicative that the score distributions for the positive and negative classes are roughly symmetric with roughly equal variances. In FAERS data, the proportion of drug-event pairs labelled as PS is typically smaller than the number not labelled PS, roughly one third, which tends to move the threshold at which sensitivity equals specificity. In the records considered here, the Prior Odds were about 0.3 depending on quarter years examined, e.g. 0.2766 for the green curve of the last 2 million records and 0.3062 for the red curve for the last available quarter year, at the time of study. Inspection of Fig. 8 shows that Prior Odds of about 0.3 will have little qualitative effect on ROC curves for which Predictability approaches 100% (top left corner) or 50% (the dotted diagonal line indicating random predictions). However, it still seems remarkable that the ROC curves are approximately symmetrical in these studies at the sample thresholds taken, and that similar approximate symmetry is in recent ADR studies by other authors for a variety of different populations, e.g., Refs [51-53], and including ROC curves where the inflection point would correspond to the 60-80% Predictable range. At least, it is

clearly sufficiently symmetrical that the Predictability score does give not only an indication of predictive power but also a useful approximate account of the ROC curve for ADR studies in practice.

4.12 Summary of Performances of Different Pharmacovigilance Models

Using Predictability because it captures estimates of the essential features of a ROC curve and ultimately expresses predictive capability that are the goals of Precision-Recall and F1 approaches, Table 2 summarizes the results in Sections 4.1 to 4.10. This is done in relation to Steps in the development of, and in some cases investigation of the possibilities of, a model with improved performance, both in alert signaling and in analytical terms. Claims of superiority must be made with caution. Gains by the TEI are necessarily small against some of the best performing standard methods simply because these methods (applied in the way specified in the Table) approach 100%, and even here a standard deviation applies, including dispersion of performance due to choice of drugs, as the last entry for step 4 exemplifies. The TEI method opens the ability to find ADRs accurately from fewer occurrences, thus alerting sooner when less damage to the population has occurred and allowing more time to assess re-labeling or changes to drug prescribing protocols. By “meaningful way” is meant based on and emergent from a fundamental theoretical position. In view of these considerations note the somewhat surprising predictions of not-C from those records showing clear results of dechallenge/rechallenge, converging rapidly with increasing data to close to 100%. It was deduced above that this was likely due to the reporter making a C or other entry based on the confidence given by such results. Importantly note that the studies otherwise carried out are not based on C versus not-C and dechallenge/rechallenge entries because records containing dechallenge and rechallenge entries represent only a small percentage of the total ADR records. This also somewhat limits the usage and value by ADR analysts.

Note that all Steps 1 to 3 are purely explorations of the TEI approach, while Steps 4 to 6 include comparisons with other standard methods, Bate’s and ROR. There is a particular interest in combining them in the same spirit as Step 4 developed the notion of Concern Classes (based on labeling, withdrawal etc.) as outcome measure and, in the last two entries for Step 4, sought to combine non-redundant information in PS, SS, I and C assignments. Step 5, in particular, shows that the methods show rather similar performance at 95-97%, when the Concern Classes used as indicators of outcome (rather than prediction of, e.g., PS). Step 7 is a return to focus on the TEI method based on the studies of sex, age, and weight in Sections 4-11, combining them as suspected impacting factors in DiracSmash which could reasonably be considered as a form of machine learning. The prediction is, as in the earliest studies, using only PS as the prediction target, but as in the preceding entries above it in the Table it is using as “experimental outcome” Concept Class grouped (“binned”) as Classes 0-1 and 2-4, with which to compare a prediction of PS or not-PS. There is a creditable, though not dramatically increased, performance of 97%. It could be considered that sex, age and weight make a modest improvement since the preceding studies in the table suggest a plausible final result in the range 95-97%, and Fig. 3 appears to portend a potential final result in the upper regions of 83-100%. Fig. 3 based on the “experimental evidence” of dechallenge/rechallenge also hints that the 97% may be an upper limit due to inevitable errors in PS, SS, I and C assignments, which are subjective. The ability to reach 97%, and perhaps marginally higher, in part depends on the use of Concern Classes used in a binary outcome manner.

Table 2. Model parameters

Step in Model Development	Section	Description of model.	Predictability analogous to accuracy = sensitivity = specificity balance point on a ROC curve.	Comments
1	4.1	Artificial data construct.	> 95%	Dummy counts used as input to study behavior with increasing data. Performance Comparable to Bayes-shrunk ROR.
2	4.5	Not-C prediction	(100%)	At circa 12 million records. May be subjectively biased entries impacted by the direct evidence of dechallenge /rechallenge entries.
2	4.5	PS prediction	83%	At circa 12 million records.
2	4.5	SS prediction	50%	At circa 12 million records.
2	4.5	I prediction	63%	At circa 12 million records.
3	4.7	Based on LR for PS only	63-89%	Range over drugs studied
3	4.7	Based on assignment of LR for PS to Concern Classes 0-1, 2-4, threshold 2	88%	Predictability estimated as probability of correct assignment
3	4.7	Combined PS-SS-I-C model Type 2 assignment to Concern Classes 0-1, 2-4, threshold 2	94%	Predictability estimated as probability of correct assignment
3	4.7	Combined PS-SS-I-C model Type 2 assignment to Concern Classes 0-1, 2-4, threshold 2	93%	Predictability estimated as probability of correct assignment
4	4.7	Literature review PRR and ROR	70-90%	Estimated from sensitivity and specificity 60-90% 70-95%
4	4.7	Literature Review EBGM	70-85%	Estimated from sensitivity and specificity 50-70%, 85-95%
4	4.7	Literature Review Bate's IC	80-85%	Estimated from sensitivity and specificity 50-80%,85-98%
4	4.7	Present study PRR and ROR	70-90%	Best efforts on common ranges of parameters and thresholds.
4	4.7	Present study EBGM	60-80%	Best efforts on common ranges of parameters and thresholds.

4	4.7	Present study Bate's IC	70-90%	Best efforts on common ranges of parameters and thresholds.
4	4.7	As above, but using only present TEI method based on Type 2 combination scores	96%	See reservations in Section 4.7. Method calculated to work on the data used.
4	4.7	Applying the same procedures as above to 5 random selections of 50% of records partitioned into tests sets/	93-97%	Some concerns of the above still apply as method was developed for this block of data and a limited set of drugs in Table 1 where marketed.
5	4.8	All methods but Bate's mild signal threshold. Based on assignment to 2 groups of Concern Classes 0-1 and 2-4, using threshold 1	95%	Methods are applied separately and Predictability is based on the dispersion of results between the upper and lower boundaries of the envelope.
5	4.8	All methods typical Bate's medium signal threshold. Based on assignment 2 groups of Concern Classes 0-1 and 2-4 using threshold 2	95%	As above.
6	4.9	Refinement of method step 5 to distinguish strong from milder signals. Involves confining envelope to upper and lower limits of Concern Classes 1,2 and 4.	95%	Above assignments into Concern Classes 1-2 and 4 based on minimum thresholds are not affected
7	4.10-4.11	LR for PS from the TEI method applied to sex, age and weight ranges predicting assignment of drugs assigned to 2 groups of Concern Classes 0-1 and 2-4	97%	Preliminary study only based on drugs in Table 1 and their Concern Classes.

Table 2. Predictability as the balance point accuracy = sensitivity = specificity for the 7 main steps in the present study. While there is a rough trend in that Predictability improves with step number, these are attempts to improve predictive capability, the effect of which may only become manifest in a later step. In step 2, simple use of primary Suspect PS gives 83% which is a reasonable first indication.

5 Discussion and Conclusions

5.1 General Observations

These studies have ultimately been investigations to explore the field of pharmacovigilance as ADR analytics, identifying data requirements, appropriate metrics (measures), and methods of analysis and interpretation. This has provided a beneficial excuse to provide a review and to serve

as a position paper. The position being taken is regarding the merits of a Theory of Expected Information that has spanned many decades [28,32,44] but for which popularity has been largely confined to bioinformatics. However, in contrast to earlier papers, the TEI in the present paper has taken on a simpler face. With direct extension to stratification/ segmentation studies in mind, Likelihood Ratio LR or Predictability $\text{Pred} = \text{LR}/(1+\text{LR})$ have been used. It remains that the LR is in this project generated as the exponential of an information measure based on zeta functions, characteristic of the TEI approach. Nonetheless, it is also true that while a formal TEI approach to combining PS, SS, I, and C was presented in Eqn. 23, focus has been on simple empirical methods of combining predictions of them (Eqns. 27-29). Primarily this is because they provide a means of using LRs for predictions of PS, SS, I, and C however they were obtained, by TEI or classically. But also, like conditional probabilities such as $P(A|B,C)$ and $P(B,C|A)$, Predictive Odds and Likelihood Ratios represent adjoint values carried on Q-UEL tags. These numbers are key players in the Q-UEL language. They enable the Q-UEL tags to be used directly in the construction of elaborate bidirectional inference nets where probabilities used are guaranteed coherent by Bays' Rule [34,37,45,46]. Bidirectional General Graphs including cyclic paths solvable without iteration are possible, and there is no longer any constraint to, for example, Directed Acyclic Graphs as in Bayes Nets [37]. They also enable automatic construction of very large odds inference nets [45,46]. All these are essentially "Glass Box" Large Probability Models rich in interactions and with fewer independency assumptions.

A beneficial consequence of the above is that hundreds, thousands, or millions of tags as reusable knowledge elements contain statistical information that can be used individually in the manner of evidence-based medicine measures, or collectively analyzed or used in prediction. Because of the TEI used and the properties of the partially summated zeta functions, tags with probabilities etc. required for predictions can be obtained and are meaningful even for single events. It is also possible to generate tags for joint events that that should have been seen based on prior expectation, but were not seen because they were negatively associated.

As far as a proposed method in ADR analytics is concerned, Eqn. 28 used in column 8 of Table 2 is a recommended combination approach because it uses more of the abundantly represented information (PS, SS, I, and C), and captures some of the intent of standard methods. Also, a method based on the combination with those standard methods would have merits, albeit more qualitatively, as illustrated by Fig. 4. It finally emerged however, that the simple LR method for PS predicted by DiracMiner [45,46] using TEI including sex, age and weight ranges emerged as consistently best as shown in the last row of Table 2. Though it may be criticized for using only PS as prediction target, it performed well when, as for other methods above, its performance assessed when applied to Concern Classes 0-1 and 2-4 for the drugs in Table 1. That is, if a PS was predicted, it was a true positive if the drug belonged to Concern Classes 2-4, and if not-PS was predicted, true negative if the drug belonged to Concern Class 0-1. Note that, although there were only 34 drugs involved in this case, which is a limitation, there is as usual a large number of records reporting PS, SS, C, or I for each drug. It remains a combination method in the sense that the Concern Classes were tested and recalibrated against other and standard methods notably regarding the treatment of Concern Class 3 described at the end of Section 4.9. Its apparent success rests on the correlation between PS reports and Concern Classes, which may vary with source data. Unlike some comparisons in the present paper, comparisons here were confined to analysis of records referencing the 34 drugs of Table 1, a useful but not large number. Any claim that this approach may be best can only be tentative for the above reasons, and as discussed in Sections 5.2 and 5.3, as follows.

5.2 Difficulties in Comparing Different Methods

As emphasized already, there are considerable difficulties in comparing previous and current established methods. Not least, the aims differ. There is the question of where on the following spectrum a method sits, from (1) alert/signal generation (ROR / PRR / IC / EBGm) signaling that something looks unusual, to (2) statistical adjustment, typically involving logistic regression (with covariates, interactions) exploring whether the signal is real, i.e., survives adjustment. That spectrum further extends to (3) epidemiologic confirmation involving health insurance claims data, medical records, cohort studies, and case control studies seeking to prove or disprove that the drug concerned is causal of the ADR. The new methods explored here essentially spans the first two; the methods of author BR are likely to be appropriate to the third are described elsewhere [33,43]. FAERS and WHO (VigiBase) core methods compared below are primarily of the first type and, in interests of fairness, the comparisons should reflect that purpose, but theoretically there is no clear boundary between considering whether a deviation from expectation is worth reporting, and trying to improve upon that, and the second step of exploring whether it is in some sense real and serious. Indeed, the question of whether it is real and serious also clearly overlaps with the question of the way and extent of it being causal.

Even methods that sit at approximately the same point on the spectrum can emphasize impact of different situations in data, and methods may agree on the general aims, but differ the fine details of exactly what is to be predicted. In this paper, prediction of not-C, PS, SS, I, and C were considered. Each established method that takes account of these kinds of assignments is at least partly tailored to the particular choice. Ideally, one should also compare methods for all the different predicted states that different methods tend to employ, and some methods do better at certain selections of prediction target than others. But often, methods address different combinations of the above targets. Testing on PS-only data is strictest, with higher specificity. Combined PS + SS data give a broader signal capture. All together (PS + SS + I + C) gives high sensitivity. The choice of including or excluding C changes the signal strength and even whether a signal appears at all. In principle, the impact of such uses should be compared for each drug at each stage of its history, and for each population. The WHO system based on Bate considers the same four states as the FDA FAERS database, but while their use in the FAERS case is mainly for data stratification and filtering (not as a causal model) the WHO approach arguably emphasizes them more as input metadata for the pharmacovigilance databases, not AI or statistical parameters themselves.

Ultimate performance criteria should involve practical industry outcomes such as strong cautionary labeling, black box warnings, and withdrawals. These are the kind of criteria used in Section 4.6. Here, matters are of course still complicated by the important consideration that for patients with rare diseases or circumstances the benefits may outweigh the risks or disadvantages. Also, they are as reported at the time of the study, a concern that can change, and the severity of labels may not always be interpreted correctly.

Comparisons even with established methods are further complicated by reproducibility concerns. In both FDA-style systems and the WHO Bate/BCPNN approach, there are data learning or parameter fitting operations. Students of the methods often consider these as machine learning and so inherent some of the “black box” character of much AI. Analysts of the FAERS database frequently use logistic regression which is often considered one of the simplest forms of machine learning, but cross-terms (interaction terms) between contributions are not treated in a single consistent way. In both FDA-style systems and the Bate/BCPNN approach, older data can be used to differing extents influence how new data is

interpreted and can change when repeated some weeks later. As more reports are added to the database, quantities such as the expected and observed frequencies and the Bayesian posterior distributions are updated. In that sense, the model does learn from the data, and so its effective parameters change with the data. Because EBGM/MGPS uses Bayesian shrinkage, hyperparameters are estimated from the entire database and are updated as more data become available. Thus, some parameters are data dependent and data changes in time. Also, they deliberately do not avoid the complication that, for patients with rare diseases or circumstances, the benefits may outweigh the risks or disadvantages.

5.3 The Need to Study a Broad Spectrum of Drug-Event Pairs

The industry well understands the need to test new methods on many drugs. Remarkably, however, many studies have been applied primarily to a single drug, e.g. Ref [54]. This one-drug approach is by no means rare in attempting to improve methodology because proper comparisons are difficult and time consuming. The drug considered in Ref [54] was Atorvastatin, and only two types of ADR symptoms were considered. Such studies are valuable in order to make certain points, but the obvious danger is that such a study or method may not represent the range of situations encountered in pharmacovigilance. In contrast, the study in Section 4.6 of the present paper considered 34 drugs essentially selected at random except that they covered a broad and reasonably distributed range from very safe to dangerous by broad accepted criteria (Table 1). The results obtained in that study soon made it apparent that one can inadvertently pick a single drug that makes any one method look very good and the rest look poor. The range of situations encountered in pharmacovigilance includes different drugs having very different reporting patterns, numbers of reports, indications, co-medications, and ADR profiles. A method that performs well for a heavily reported drug may perform poorly for a rarely reported drug. For just one drug, the choice reported in a paper might just be lucky as far as good prediction is concerned, and there is the danger that a new method being developed might have been inadvertently partly refined with respect to that single drug case. A researcher should at least develop a method based on one drug and test it on several others. But even that would be complicated by the fact that methods belonging to classes represented by PRR, ROR, IC, and EBGM behave differently for different data, notably when counts are sparse versus abundant. Further, there is a risk of selection bias: choosing a well-known drug with established ADRs may favor some methods while choosing a drug with few reports may favor methods designed for sparse data. The results obtained in the present study endorse best practice that researchers should include several drugs with high reporting volume, several with low reporting volume, drugs from different therapeutic classes, and multiple known drug-ADR pairs (positive controls). Admittedly, these are matters of degree, scope, and to some extent of opinion, but intuitively and fundamentally it seems unlikely that conclusions drawn from one drug can be generalized to other drugs, therapeutic classes, or reporting environments. A strong study design should include multiple drugs from different therapeutic classes, drugs with high, medium, and low reporting frequencies, with both well-established and emerging safety signals, and ideally, several known positive controls and negative controls.

5.4 Future Work

Use of LR, Pred, and ROR* expressed in terms of partially summated zeta functions $\zeta(s,n)$, provides a solid, formal, non-empirical “*ab initio*” basis that should be extended to large studies involving more drugs, and direct use as improved signaling methods in which other values of s , used in analysis of Black Swan events [38, 39], are explored. Not least, the methods should be extended to exploit for pharmacovigilance the rich information in general medical health records, especially longitudinal (cradle to grave) health records. The Theory of Expected Information has already been

used to analyze large numbers of medical records for clinical decision support [45,46], retrospective cohort studies [43] and clinical trials [44]. The hope is that the richer detail in medical records will help provide diverse evidence, including counterfactual evidence, to show that a suspect drug is truly causal of an ADR, by introducing many more impacting demographic and clinical factors than just sex, age and weight as considered here. The new paths described here are well suited to pharmaceutical business intelligence, but the main limitation for speedy analysis is arguably that “Correlation does not imply causation” [41]. Currently, causality is usually addressed outside the signal metrics, e.g., the Naranjo scale [42], based on answers to questions such as “Did the adverse event appear after the drug was given?”. This is a manual, expert-enabled process (as are many of the other methods). As an enhancement, essentially Bayesian methods underlying the TEI can be automated to help support Naranjo and other manual findings. Regarding that and more immediate next steps concerning causality, note that the TEI is only one pillar of the larger system based on the Q-UEL language and the Hyperbolic Dirac (inference) Net for interoperability and probabilistic knowledge management in medicine (e.g. Refs [37,44-46]). It is well-constructed to explore bidirectional conditionality as discussed above, and hence causality. While external knowledge, not just probability or odds calculation, is required to establish causality [41] (despite claims by some Bayes Net users), a first minimal requirement is to be able to see in both directions of conditionality in an integrated way that means conforming to Bayes’ Rule, one of which directions will point to cause (etiology) and the other to outcome. Reinterpreting that approach to causality as extensions to the TEI brings ADR analysis and prediction closer to measures of Evidence Based Medicine and the current predictive goals of AI. Such a convergence could be important because these two disciplines are not, in the present state of ADR analytics, usually considered a likely marriage.

Such a study using general medical records was originally a later section of this manuscript, but this added greatly to length and essentially only extended a study described elsewhere [43]. That involved different combinations of pharmaceutical and surgical treatments for bleeding peptic ulcer in which initial and final clinical laboratory (“bloodwork”) values did or did not show a desirable outcome to treatment. Results were promising. In those preliminary studies, A high LR corresponding to Predictability ≥ 65 was consistent with a high ROR* as reported in Refs [43,44] except in two cases where there was no ADR evidence as seen in the final laboratory measures: electrolytes finished as normal, and LDH went from high to normal. This kind of observation would be more reliable and relevant to the present project when direct data reporting an adverse drug reaction is available within the medical record and clearly relating to the course of treatment. Hopefully, such data will be soon available to the present authors and collaborators.

Acknowledgements

The authors are grateful for input from Srinidhi Boray of Engine Inc. as to the exploration into the state of the art and reference to information sources.

Conflicts of Interest

The principal purpose of the present paper is to illustrate theoretical principles and perform analyses to obtain results of interest. It used the software of startup Engine Inc. This software has previously been described in many peer-reviewed publications, several key examples of which are cited in the present paper, but the project described here is a new application. Barry Robson has performed tasks for Engine in the role of consultant. These consulting tasks performed to date did not primarily involve

the product of interest and were essentially theoretical. Both the authors have potential commercial interests in Ingene Inc. as cofounder and CEO, respectively. B. Robson is also Chief Executive Office of The Dirac Foundation which is non-commercial and in which he has no financial interests. This foundation has the charter to promote the ideas of Nobel Laureate Professor Paul A. M. Dirac for the benefit of human and animal medicine, and the information in the present paper is presented in that spirit.

Funding

There was no external funding for this project.

Ethics consent and permission

This study did not involve human beings. The studies from publicly available US FDA data are as described below. Some data was originally from deidentified patient data and involved only statistical summaries over many hundreds of thousands of records, but these studies were removed for brevity (to be described in detail elsewhere) except for a brief summary in Section 5.4 represented as preliminarily observations for future study. No permission was required for these studies. While the software used is proprietary to Ingene Inc., and used with the permission of Ingene Inc., accounts of mathematics and algorithms used are given in Refs [7,8,25,30-34,37,43-46,50] and references cited therein, and in earlier papers also referenced here and applied to data mining and prediction in bioinformatics, e.g. Refs [26-30].

Data availability

This study used the publicly available US FDA Adverse Event Reporting System (FAERS) Quarterly Data Extract Files, (2025) <https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html>.

References

1. J. Mandel, The Statistical Analysis of Experimental Data, Interscience Publishers, John Wiley and Sons, NY, (1964)
2. N. Heard, An Introduction to Bayesian Inference, Methods and Computation, Springer, (2021), DOI: 10.1007/978-3-030-82808-0.
3. <https://www.fda.gov/drugs/fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard> (last accessed Dec 9, 2025).
4. T. Sakaeda, A. Tamon, K. Kadoyama, and Y. Okuno, Data mining of the public version of the FDA Adverse Event Reporting System. International Journal of Medical Science, 10, 7, 796-803, (2013). DOI:10.7150/ijms.8.332.
5. <https://retractionwatch.com/2025/09/16/exclusive-journal-bans-drug-safety-database-papers-as-they-flood-the-literature/> (last accessed Dec 9, 2025).
6. FDA, Sentinel System 5-Year Strategy 2019-2023, <https://www.fda.gov/media/120333/download>. (last accessed Dec 8, 2025).
7. B. Robson, T. P. Caruso, U. G. J. Balis, Suggestions for a web based universal exchange and inference language for medicine. Continuity of patient care with PCAST disaggregation, Computers in Biology and Medicine 56, 51-66, (2015). DOI:10.1016/j.compbiomed.2014.10.022.
8. B Robson and O.K. Baek, An ontology for very large numbers of longitudinal health records to facilitate data mining and machine learning, Informatics in Medicine Unlocked 38, 101204, (2023). DOI: 10.1016/j.imu.2023.101204.

9. M. Bright, E. Kontsioti, M. Pirmohamed et al., MCSG: A Method for Simultaneous Disproportionality Analysis and Background Rate Estimation in Large Pharmacovigilance Databases, *Drug Safety*, (2025). DOI:10.1007/s40264-025-01632-8
10. N. N. Taleb, *The Black Swan: The Impact of the Highly Improbable* (2nd ed.). London: Penguin (2007).
11. W. D. Nordhaus, *An Analysis of the Dismal Theorem*. Cowles Foundation Discussion Paper No. 1686, (2009).
12. A. Lee and M. Cuthbert, *Adverse Drug Reactions*, Pharmaceutical Press, (2023).
13. R. D. Gibbons and A. Amatya, *Statistical Methods for Drug Safety*, Chapman & Hall/CRC Biostatistics Series, (2023).
14. T. J. Hwang, D. Carpenter, J. C. Lauffenburger, B. Wang, J. M. Franklin, and A. S. Kesselheim, Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Internal Medicine*, 176, 12, 1826-1833, (2016). DOI:10.1001/jamainternmed.2016.6008.
15. D. B. Fogel, Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, 11, 156-164, (2018). DOI: 10.1016/j.conctc.2018.08.001.
16. https://en.wikipedia.org/wiki/List_of_disasters_by_cost, (last accessed April 18, 2023).
17. C. R. Buncher and J. Y. Tsay (Eds), *Statistics In the Pharmaceutical Industry*, CRC Press, (2023).
18. S. R. Evans, Common statistical concerns in clinical trials. *J. Exp Stroke Transl Med*. 3,1, 1-7, (2010). DOI:10.6030/1939-067x-3.1.1.
19. D. W. Light, J. Lexchin, J. J. Darrow, Institutional Corruption of Pharmaceuticals and the Myth of Safe and Effective Drugs, *Journal of Law, Medicine and Ethics*, 14, 3, 590-610, (2013). DOI: 10.1111/jlme.12068.
20. K. Trzuppek and J. Johnston, Adding genetics to the clinical trial ecosystem, *MedCity News*, December, (2018), <https://medcitynews.com/2018/12/adding-genetics-to-the-clinical-trial-ecosystem/>(last accessed Aug 18, 2025).
21. D. J. Finney, Systemic signaling of adverse reactions to drugs, *Methods Inf Med*, 1974 13,1, 1-10, (1974). DOI: 10.1055/s-0038-1636131.
22. S. J Evans, P. C. Waller, and S. Davis, Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology, Drug Safety*, 10.6,483-6, (2001). DOI:10.1002/pds.677.
23. A. Bate, *The Use of a Bayesian Confidence Propagation Neural Network in Pharmacovigilance*, Doctoral Dissertation (Thesis), Umeå university, (2003). <https://www.diva-portal.org/smash/get/diva2:144650/FULLTEXT01.pdf> (last accessed 26/08/2025).
24. M. Lindquist, M., *The WHO Programme for International Drug Monitoring: The Present and Future*, in *Electronic Communication Technologies*, M. Mitchard, Editor, 527-549, (1998).
25. B. Robson, Towards new tools for pharmacoepidemiology, *Advances in Pharmacoepidemiology and Drug Safety*, 1-6, (2013). <https://www.walshmedicalmedia.com/open-access/towards-new-tools-for-pharmacoepidemiology-2167-1052.1000123.pdf>. DOI: 10.4172/2167-1052.1000123.
26. R. H. Pain and B. Robson, Analysis of the Code Relating Sequence to Conformation in Globular Proteins, *Nature*, 227, 62-63 (1970). DOI:10.1038/227062a0.
27. B. Robson and R.H. Pain, Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions, *Journal of Molecular Biology* 58, 237-257, (1971).
28. B. Robson, Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Theory and Application of Expected Information, *Biochemical Journal*, (1974) 141, 853-867 (1974).
29. J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins", *J. Mol. Biol.* 120, 97-120, (1978). DOI: 10.1016/0022-2836(71)90243-9.
30. B. Robson, Testing machine learning techniques for general application by using protein secondary structure prediction. A brief survey with studies of pitfalls and benefits using a simple progressive

- learning approach, *Computers in Biology and Medicine*, 138, (2021). DOI:10.1016/j.combiomed.2021.104883.
31. B. Robson, The Dragon on the Gold: Myths and Realities for Data Mining in Biotechnology using Digital and Molecular Libraries", *J. Proteome Res. (Am. Chem. Soc.)* 3 (6), 1113 - 9, (2004). DOI: 10.1021/pr0499242.
 32. B. Robson, Clinical and Pharmacogenomic Data Mining: 3. Zeta Theory as a General Tactic for Clinical Bioinformatics, *J. Proteome Research (Am. Chem. Soc.)* 4(2); 445-455 (2005). DOI: 10.1021/pr049800p.
 33. I. M. Mullins, M. S. Siadat, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, W. A. Knaus, Data mining and clinical data repositories: Insights from a 667,000 patient dataset, *Computers in Biology and Medicine*, 36, 12, 1351-1377, (2006). DOI:10.1016/j.combiomed.2005.08.003.
 34. B. Robson, B. and S. Boray, Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities and inference in data mining of clinical data repositories, *Computers in Biology and Medicine*, 66, 82-102 (2015). DOI:10.1016/j.combiomed.2015.07.015.
 35. J. Pearl, *Probabilistic Reasoning in intelligent systems*, San Mateo, Morgan Kaufmann (1988).
 36. J. Pearl, *The Book of Why*, Allen Lane, Penguin Books (2018).
 37. Robson, B., Bidirectional General Graphs for inference. Principles and implications for medicine, *Computers in Biology and Medicine*, 10, 382-399 (2019). DOI:10.1016/j.combiomed.2019.04.005.
 38. G. De Marzo, A. Gabrielli, A. Zaccaria, and L. Pietronero, Quantifying the Unexpected: a scientific approach to Black Swans, <https://arxiv.org/pdf/2201.12042.pdf>, (2002). DOI:10.1103/PhysRevResearch.4.033079
 39. G. K. Zipf, The Unity of Nature, Least-Action, and Natural Social Science, *Sociometry*. 5, 48-62, (1942). DOI:10.2307/2784953.
 40. J. S. Almenoff, K. K. LaCroix, N. A. Yuen, D. Fra, and W. DuMouchel, Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department, *Drug Safety*. 2006;29, 10, 875-87, (2006). DOI:10.2165/00002018-200629100-00005.
 41. R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd (Edinburgh), (1925).
 42. C. A. Naranjo, U. Busto, E. M. Sellers, P. Sandor, P. I Ruiz, E.A. Roberts, E. Janeczek, C. Domecq, D. J. Greenblatt. A method for estimating the probability of adverse drug reactions, *Clinical Pharmacology and Therapeutics*. 30,2,239-245,(1981). DOI: 10.1038/clpt.1981.169.
 43. B. Robson and O.K. Baek, Glass box machine learning for retrospective cohort studies using many patient records. The complex example of bleeding peptide ulcer", *Computers in Biology and Medicine*, 108085, (2024). DOI: 10.1016/j.combiomed.2024.108085.
 44. B. Robson and O.K. Baek, Use of a theory of expected information for sparse data and adverse events in clinical trials and other biomedical studies", *Information Sciences*, Volume 680, October, 121027, (2024). DOI: 10.1016/j.ins.2024.121027.
 45. B. Robson, B. and S. Boray, Studies in the Extensively Automatic Construction of Large Odds-Based Inference Networks from Structured Data. Examples from Medical, Bioinformatics, and Health Insurance Claims Data, *Computers in Biology and Medicine*, 95:147-166 (2018). DOI:10.1016/j.combiomed.2018.02.008.
 46. B. Robson, S. Boray, and J. Weisman, Mining real-world high dimensional structured data in medicine and its use in decision support. Some different perspectives on unknowns, interdependency, and distinguishability, *Computers in biology and Medicine*, 141,105118,(2021). DOI:10.1016/j.combiomed.2021.105118.
 47. FDA, FDA Adverse Event Reporting System (FAERS) Quarterly Data Extract Files, (2025) <https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html> (Last accessed Aug 19, 2025).
 48. ASC_NTS.DOC file for the quarterly data extract (QDE) from the FDA Adverse Event Reporting System (FAERS) U.S. Food and Drug Administration (FDA) Center for Drug Evaluation and Research

- (CDER) Office Of Surveillance And Epidemiology (OSE), last revised: January 2025. (Last accessed Aug 19, 2025).
49. C. T. Nakas, L. E. Bantis, and C. A. Gatsonis, ROC Analysis for Classification and Prediction in Practice, Chapman & Hall, CRC Press, (2023).
 50. B. Robson, R. Cooper, Glass Box and Black Box Machine Learning Approaches to Exploit Compositional Descriptors of Molecules in Drug Discovery and Aid the Medicinal Chemist, ChemMedChem , 19,18, e202400169 (2024). DOI:10.1002/cmdc.202400169.
 51. A. Farnoush, Z. Sedighi-Maman, B. Rasoolian, J.J. Heath, and B. Fallah, Prediction of adverse drug reactions using demographic and non-clinical drug characteristics in FAERS data, Scientific Reports, 14, 1, 23746, (2024).
 52. J. Frydenlund et al., Adverse drug reactions and events in an Ageing PopulaTion risk Prediction (ADAPTIP) tool: the development and validation of a model for predicting adverse drug reactions and events in older patients, European Geriatric Medicine, 16, 573-581, (2025)
 53. Y. Tian, J. Yi, K. Li, J. Pen, Y. Deng, and D. Cao, Machine learning prediction of pediatric adverse drug reactions using consensus-derived scarce data, Commun Chem 9, 59 (2026).
 54. Y. Ding, M. Markatou and R. Ball, An evaluation of statistical approaches to postmarketing surveillance, Statistics in Medicine, 39, 845-874,(2020).