CrossMark

SPECIAL ARTICLE

Good Signal Detection Practices: Evidence from IMI PROTECT

Antoni F. Z. Wisniewski[1] · Andrew Bate[2] · Cedric Bousquet[3,4] · Andreas Brueckner[5] ·
Gianmario Candore[6] · Kristina Juhlin[7] · Miguel A. Macia-Martinez[8] ·
Katrin Manlik[9] · Naashika Quarcoo[10] · Suzie Seabroke[11] · Jim Slattery[6] ·
Harry Southworth[12] · Bharat Thakrar[13] · Phil Tregunno[11] · Lionel Van Holle[14] ·
Michael Kayser[15] · G. Niklas Norén[7]

Abstract Over a period of 5 years, the Innovative
Medicines Initiative PROTECT (Pharmacoepidemiological
Research on Outcomes of Therapeutics by a European
ConsorTium) project has addressed key research questions
relevant to the science of safety signal detection. The
results of studies conducted into quantitative signal detec-
tion in spontaneous reporting, clinical trial and electronic
health records databases are summarised and 39 recom-
mendations have been formulated, many based on com-
parative analyses across a range of databases (e.g.
regulatory, pharmaceutical company). The

✉ Antoni F. Z. Wisniewski
  antoni.wisniewski@astrazeneca.com

1   AstraZeneca, Macclesfield, UK

2   Pfizer, Walton-on-the-Hill, Surrey, UK

3   INSERM, UMR_S1142, LIMICS, Paris, France

4   Department of Public Health and Medical Informatics, CHU
    University Hospital of Saint Etienne, Saint-Étienne, France

5   Novartis, Basel, Switzerland

6   European Medicines Agency, London, UK

7   Uppsala Monitoring Centre, Uppsala, Sweden

8   Agencia Española de Medicamentos y Productos Sanitarios,
    Madrid, Spain

9   Bayer Pharma AG, Berlin, Germany

10 GlaxoSmithKline, London, UK

11 Medicines and Healthcare Products Regulatory Agency,
    London, UK

12 Data Clarity Consulting, Stockport, UK

13 F. Hoffmann-La Roche, Basel, Switzerland

14 GlaxoSmithKline, Wavre, Belgium

15 Bayer Pharma AG, Wuppertal, Germany

recommendations point to pragmatic steps that those
working in the pharmacovigilance community can take to
improve signal detection practices, whether in a national or
international agency or in a pharmaceutical company set-
ting. PROTECT has also pointed to areas of potentially
fruitful future research and some areas where further effort
is likely to yield less.

1 Introduction

The opportunities for effective signal detection in large
databases have improved substantially since the early days
of pharmacovigilance. In those early days, much effort
necessarily focussed on manual clinical review of incom-
ing reports—often in the form of handwritten or typed
reports of suspected adverse drug reactions (ADRs) by
pharmacovigilance experts. Adverse event reports, or sets
of reports, in some way triggering the suspicion of a clin-
ical reviewer would be further investigated by the reviewer
and in some instances go on to become signals with some
of these signals leading to further actions. While sophisti-
cated approaches to statistical signal detection had been
proposed and even tested in a limited methodological or
theoretical context [1–5], routine prospective screening
using tools and automated systems was a mere pipe dream.
Several decades later, systematic screening of adverse
event reports is not just a reality; it is today the de facto
standard in large datasets [6]. However, as adverse event
reports are exchanged electronically around the world, in
the tens of thousands on a daily basis, it is well accepted
that our capabilities for signal detection are far from per-
fect and should be improved. In addition, we are seeing a
significant shift of focus, beyond adverse event reports and
prescription event monitoring systems, on to the use of

△ Adis

longitudinal observational databases and clinical trials for signal detection. Although the former have been used extensively in the past for epidemiological purposes, their potential use for signal detection is novel and new approaches for identifying safety signals in clinical study data beyond basic comparisons of adverse event frequencies have been slow to gain widespread use.

CIOMS VIII [7] was convened in part to address the widespread and growing interest in quantitative analysis of spontaneous reports and particularly the desire for guidelines around the appropriate use of such quantitative approaches. There was also a need to contextualise the role and use of such approaches, including disproportionality, within a holistic signal management perspective. CIOMS VIII was able to achieve these ambitions but also highlighted several areas of signal detection in spontaneous report and observational data and in the use of terminologies that needed more research. PROTECT was set up to address these, and a number of other important topics, in signal detection.

The following recommendations[1] relate to research conducted into quantitative signal detection in spontaneous reporting, clinical trial and electronic health records databases conducted under the auspices of the Innovative Medicines Initiative (IMI) PROTECT Work Package 3 (one of seven work packages) between September 2009 and February 2015. The PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium; http://www.imi-protect.eu) was a public–private partnership co-ordinated by the European Medicines Agency (EMA) and received support from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement No. 115004. Full details of IMI PROTECT, including specific information about Work Package 3 and a detailed technical report supporting the recommendations can be found at http://www.imi-protect.eu/about.shtml.

In attempting to advance the science of safety quantitative signal detection and with the overall goal of increasing the efficiency of quantitative signal detection practices, the scope of IMI PROTECT Work Package 3 has been ambitious, with 12 separate work streams covering a wide range of research questions and undertaken between September 2009 and February 2015. To convert the insights gained from these efforts into meaningful and executable outputs, the following recommendations have been developed: in total, 39 separate recommendations have been formulated, many based on comparative

analyses across a range of databases (e.g. regulatory, pharmaceutical company) and in several data sources (spontaneous reporting, clinical trials and electronic patient records). A further 25 recommendations for future research are also offered. It is acknowledged that signal detection relies on quantitative and qualitative elements and it should be noted that the focus of Work Package 3 was primarily concerning the former. The majority of recommendations are based on the outcomes of the signal detection research conducted under the auspices of IMI PROTECT Work Package 3. However, in attempting to provide holistic and unambiguous messages of how the work can impact good signal detection practices, some reference to recommendations beyond the strict scope of PROTECT is necessary.

2 Précis of the Research and Recommendations

The studies undertaken though IMI PROTECT, which provide the evidence base for each set of recommendations that follow, are summarised here. For a full description of the studies, refer to the cited original publication or technical report.

2.1 Timeliness of Quantitative Signal Detection Using MedDRA® Terms and Groupings

Different terms in standard medical terminologies can be used to describe the same suspected ADR. Many organisations rely on disproportionality analysis for first-pass screening of large collections of individual case safety reports, in which the observed rate of a drug and adverse reaction reported together is compared with an expected value based on their relative frequencies reported individually in the spontaneous reporting database. Confidence intervals or statistical significance tests are used to provide some protection against spurious associations and a certain number of reports are generally required before an association can be detected. By grouping together related medical terms for the purpose of analysis, the observed count will increase, but so too may the corresponding expected value. It is not known whether lumping or splitting is preferable for timely quantitative signal detection. A previous study reported lower sensitivity but higher positive predictive value for MedDRA® groupings than for preferred terms (PTs) using cumulative data, but did not evaluate the timeliness of statistical signalling [8].

The study of Hill et al. [9] sought to determine to what extent the use of standard MedDRA®[2] groupings could expedite the detection of disproportionate reporting patterns for historical safety signals, relative to analysis by

---

[1] The non-binding recommendations presented in this report represent the views of the authors and do not represent the views or policies of the authors' respective affiliations (unless by coincidence), even if employees of those organisations at the time of preparing this paper.

---

[2] MedDRA is *Medical Dictionary for Regulatory Activities*.

individual MedDRA® PTs, separately, as is common practice. Analyses were performed in the World Health Organization (WHO) Global Individual Case Safety Reports Database, VigiBase® as of 5 February, 2010. The scope of the study was restricted to 13 medical concepts identified as having medium to high probability of being drug related [10]. The ADRs consisted of 43 historical labelling changes by the EMA related to one of the 13 medical concepts, derived from a previously published reference set [11]. Each labelling change had an associated index date indicating when the EMA first became aware of the potential signal and initiated their investigation, which was used for reference. For each medical concept, related high-level terms (HLTs), narrow standardised MedDRA® queries (SMQs) and PTs were manually identified. The latter were selected among the PT associated with either the HLT or the SMQ for each concept. For each medical concept, separate analysis of each PT was conducted and also a joint analysis of all related PTs together as a custom group. Disproportionality analysis was based on the information component (IC), adjusted for country of origin and year of submission to VigiBase (simultaneously), through a Mantel–Haenszel-type stratification, as described previously [12]. Lower 95 % two-sided credibility interval limits of the IC (IC_{025}) were computed retrospectively for each quarter of a year from 1995 to 2010. For each level of the hierarchy, the quarter in which IC_{025} first exceeded zero was determined. For the analysis using individual PTs, this was the first quarter that an IC_{025} value related to any single PT in the group exceeded zero.

The study found no overall benefit in conducting signal detection using MedDRA® HLTs or SMQs compared with using PTs. Some relatively minor gain in time to signalling was seen when closely related (in a clinical sense) ADR terms where grouped together and this should be explored further in future studies.

Recommendation	Rationale
For overall timeliness in quantitative signal detection, analysis can be performed at the MedDRA® PT level	The PROTECT study found no advantage in conducting signal detection at levels of MedDRA® above the PT level and indeed observed a net loss in timeliness of quantitative signal detection from replacing an analysis at the PT level with one at a higher level of the hierarchy [9]
Future research should evaluate the false-positive burden for signal detection at each level of the terminology	The false-positive burden was out of scope for the PROTECT study [9]

Recommendation	Rationale
Future research should evaluate tighter custom-made groupings of MedDRA® PTs for signal detection	Neither PTs nor HLTs are universally ideal for quantitative signal detection. Gains in time by aggregating PTs were observed in the PROTECT study when the terms were very similar, in a clinical sense [9]
Future research should evaluate simultaneous analysis at different levels of the terminology	Parallel analyses at different terminological levels could improve timeliness but have resource implications [9]
Future research should explore a broader range of ADRs	The PROTECT study was restricted to a selection of 13 ADR categories [9]

2.2 Use of Novel Term Groupings Generated by Knowledge Engineering Techniques

New methods based on knowledge engineering techniques have been used to support the development of new groupings of terms or new terminologies, respectively. For example, the French Common Classification of Clinical Procedures (CCAM) was built using artificial intelligence tools from the European GALEN (Generalized Architecture for Languages, Encyclopaedias and Nomenclatures) project [13]. Version 11 of the *International Classification of Diseases* (ICD), which includes reference to multiple factors such as body systems, symptoms and causal agents, was developed using the experience of international experts in medical informatics [14]. Thus, knowledge engineering techniques can be used to derive novel groupings of adverse event terms based on semantic definitions of each term [15] and these groupings may provide an alternative to the standard groupings available in an adverse event terminology, such as MedDRA® HLTs or SMQs. In particular, knowledge engineering may allow for a more flexible approach to defining groups, based on the relevant dimensions for a specific topic of interest.

Two PROTECT studies [16, 17] employed a bespoke ontology (OntoADR [17, 18]) created using formal definitions of MedDRA® PTs. The formal definitions were either inherited from mapped SNOMED clinical terms or defined in semi-automatic or manual processes [18]. The semantic definition for the MedDRA® PT, 'Upper gastrointestinal haemorrhage', is illustrated as an example:

Upper gastrointestinal haemorrhage......		
Descriptor used by computer	Clinical meaning	Object in this example
hasFindingSite	Affected body structure	Upper gastrointestinal tract structure
hasAssociatedMorphology	Morphologic abnormality	Haemorrhage
interprets	Procedure undertaken	Endoscopy
hasDefinitionalManifestation	Sign, symptom or diagnosis	Melena
hasDefinitionalManifestation	Sign, symptom or diagnosis	Upper gastrointestinal haemorrhage
hasDefinitionalManifestation	Sign, symptom or diagnosis	Vomit: blood present

The following type of query can then be applied to retrieve adverse event terms related to the concept 'Upper gastrointestinal bleeding' [14]:

hasFindingSite some 'Upper gastrointestinal tract structure' AND *hasAssociatedMorphology some* 'Haemorrhage'

OR

interprets some 'Occult blood screening' AND *hasInterpretation some* 'Positive'

OR

interprets some 'Evaluation of stool specimen' AND *hasAssociatedMorphology some* 'Haemorrhage'

The concordance between groupings derived with knowledge engineering techniques and standardised event groupings, or manually derived groupings were studied and found fair concordance between the terms included in the standard and proposed groupings [16, 17, 19], as well as high concordance between the corresponding measures of disproportionality with randomly selected drugs in the FAERS[3] database [16, 17].

Recommendation	Rationale
Knowledge engineering techniques may be considered as an adjunct to the creation of custom groupings and SMQs designed for the selection and extraction of case reports in pharmacovigilance databases	The PROTECT studies show it is possible to propose relevant novel groupings when no predefined grouping is available in MedDRA® for a given safety topic (e.g. anaphylactic shock or upper gastric hemorrhages) [16, 17]
Additional research would be necessary to validate if novel groupings generated by knowledge engineering techniques can help in the design of appropriate groupings of MedDRA® PTs for use in signal detection or evaluation	Given the current state of research, the clinical accuracy of groupings generated by knowledge engineering is such that manual clinical review is still required and this still needs to be validated against existing signal detection methods [16, 17, 19]

Recommendation	Rationale
Consideration should be given to piloting the use of knowledge engineering in developing groupings in other ontologies for application to other vocabularies and their possible linkage	Given that it has been shown to be possible to generate relevant novel groupings in MedDRA®, it is reasonable to expect that it would also be possible in other ontologies, e.g. ICD10, SNOMED

2.3 Development of a Structured Database of SPC §4.8 as Reference Dataset

The information available from the Summary of Product Characteristics (SPC) is increasingly used by computer applications. For a computer program to make use of this information, it must be coded according to a well-defined and exhaustive dictionary. Probably the most commonly used dictionary for ADRs is MedDRA®. It is recommended [20] that ADRs in the SPC be listed using exact (usually preferred) terms from MedDRA® but this does not always occur in practice. Sometimes this is because the ADR could be described using multiple MedDR^A® PTs and would thus be difficult to communicate efficiently to clinical staff in this format. Of course, no such problem arises with a computer that can handle multiple terms efficiently. Thus, it is worth investing the quite considerable effort needed to convert the ADR information in the SPC for use in computer applications. PROTECT took on the task of creating a structured database of the ADR information in section 4.8 of the SPC for all European centrally authorised products using MedDRA® as the coding dictionary.

---

[3] The Food and Drug Administration Adverse Event Reporting System (FAERS) is a database that contains information on adverse event and medication error reports submitted to the US Food and Drug Administration.

The coding of the SPC was a three-stage process. First, strings of text corresponding to discrete concepts were extracted manually and an exact match to the MedDRA® hierarchical dictionary at some level was sought using simple procedures in SAS. If this first matching failed, the EMA sent the list of unmatched codes to the Uppsala Monitoring Centre (UMC) where a fuzzy matching algorithm was run to identify potential alternative matches to the strings [21]. The final step was to subject remaining codes to expert evaluation at the EMA and Bayer Pharmaceuticals.

An obvious application of machine-readable ADR information arises in pharmacovigilance signal detection systems. Such systems use ADR reports that are themselves coded in MedDRA® and hence it is easy to use the ADR dataset to determine immediately if a new signal of disproportionate reporting (SDR) [22, 23] corresponds to a known ADR. This eliminates the need for manual inspection of the SPC where the focus of the monitoring is the detection of new risks. This use has already been tested successfully at the EMA and the UMC and implemented into their corresponding signal detection process. Another potential application would be to alert doctors to a possible known ADR using the information on medications and clinical events contained in electronic health records. A further use is in research on ADRs, for example, it could be used to identify products with similar ADR profiles or to help construct reference datasets for evaluation of quantitative signal detection algorithms [21, 24]. If similar databases are constructed in other regulatory systems, they could also be used to investigate the consistency of regulatory decisions across these systems.

A difficulty with mapping textual SPCs to MedDRA® is that a given medical concept may map to several PTs. When HLTs from MedDRA® are used in the SPC this is handled easily but non-MedDRA® terms require considerable thought. No systematic process exists to build up groups of MedDRA® PTs corresponding to broader medical concepts. It is currently an incremental process that improves over time as non-standard terms are discovered and the mappings refined through repeated scrutiny, and this area needs additional work. It would also be desirable to extend the database to products authorised through national procedures.

Recommendation	Rationale
The structured database of ADRs for centrally authorised products may be used as a reference to enhance pharmacovigilance for these products. The database is available here: http://www.imi-protect.eu/adverseDrugReactions.shtml [25]	The PROTECT database has been used to provide a reference in evaluating signal detection methods and also to identify known ADRs emerging from routine signal detection activity, hence reducing unnecessary investigation. Other potential uses have been identified but not yet tested
Structured databases of ADRs and their synonyms mapped to MedDRA® should be set up to cover other products	The current database does not address the majority of products authorised under mutual recognition or national processes. Further work is required if similar benefits are to be realised in pharmacovigilance systems covering these products
A standard minimum structure should be established for all SPC ADR databases. The PROTECT database provides a useful template for this structure	To maintain the utility of databases and allow combinations across databases, a standardised core structure will be essential although the appropriate structure will depend on the intended functions of the database. Thus, a coordinated approach with wide consultation of intended users would be needed. Co-ordination of such an effort could be undertaken by a large regulatory agency or a cross-industry organisation. For a description of the database structure see: http://www.imi-protect.eu/documents/Databasestructure.pdf [26]
To facilitate signal detection, exact MedDRA® terms should be used to identify ADRs in SPC section 4.8 where feasible. When an ADR involves very large numbers of terms and requires an ad hoc name, mapping from this name to the relevant MedDRA® terms should be maintained	This is essential to facilitate the construction of machine-readable data sources that have a number of potential uses including the facilitation of signal detection. See Eudralex Vol. 2: http://ec.europa.eu/health/documents/eudralex/vol-2/index_en.htm [27]

Recommendation	Rationale
Computerised text processing to help in mapping non-standard descriptions of ADRs to MedDRA® codes should be considered both for efficiency and consistency of coding practice	The approximate matching system was used to find appropriate MedDRA® terms when non-standard terminology was used in the SPC. This was usually successful and also much more efficient than human intervention alone [28]
In setting up a database of ADRs, a programme of maintenance should be established to reflect changes to the SPC from emerging safety issues or MedDRA® version changes	Around half the ADRs listed in SPCs are added as a result of post-authorisation activities and hence the database will require continuous attention to keep it up to date
Consideration should be given to establishing the value and feasibility of having direct links between databases of SPC data and other product information sources to prevent the need for duplicate data sources, or avoid repetition in the types of data collected in the different sources	Lists of product ADRs are currently maintained by regulators and by MAHs. These may conflict either in detail or in coding conventions. Even when they agree, it is not efficient to maintain independent sources of identical data

2.4 Comparison of Disproportionality Analysis Methods Within and Across Spontaneous Report Databases

Most pharmacovigilance departments maintain a system to identify ADRs through analysis of spontaneous reports. The majority of statistical methods used employ a disproportionality statistic calculated for each drug-event combination in the dataset and a signal detection algorithm that consists of a set of conditions that the disproportionality statistic and, possibly, other statistics calculated for the drug-event combination must satisfy for a SDR to be identified. The nature of the spontaneous report databases in terms of size and drug diversity varies between operators and it is unclear whether any signal detection algorithm can be expected to provide good performance in a wide range of environments. A number of different disproportionality statistics are in use [12, 29–31], but they are conceptually very similar [32–34].

The study of Candore et al. [35] compared the performance of a number of commonly used signal detection algorithms used across a range of spontaneous report databases at national and international pharmacovigilance organisations and individual pharmaceutical companies. A set of 220 products was chosen and a reference set of ADRs was compiled based on SPC and company core data sheets. Among four companies, one national agency and two international spontaneous report databases, 15 quantitative signal detection algorithms based on five disproportionality statistics were tested using a subset of products that fell within the pharmacovigilance responsibilities of

the respective database owners. Signals of disproportionate reporting were identified at monthly intervals and classified as true positives if they corresponded to an entry in the reference set. To measure the (algorithm's) performance, these results were summarised as sensitivity and precision (specifically, the positive predictive value) for each algorithm in each database. Time to signalling was also investigated, as early detection is an important contributory factor to effective pharmacovigilance.

Different algorithms gave very different levels of signal detection performance across all spontaneous report databases tested. However, increases in sensitivity were generally associated with a decrease in precision and no method clearly dominated all others. The performance is strongly dependent on the thresholds and other rules based on the disproportionality statistics that define a statistical signal. However, the different disproportionality statistics did not themselves influence the achievable performance: the choice of signal detection algorithm was much more important than the choice of disproportionality statistic. Absolute performance of the same algorithm might be very different between one spontaneous report database and another but the relative performance of two algorithms was generally similar in different databases. Over the lifetime of a product, there is a reduction in precision of any quantitative signal detection algorithm.

The changes in sensitivity and precision obtainable by replacing one quantitative signal detection algorithm with another are predictable. However, the absolute performance of a method is specific to the spontaneous report database and is best assessed directly on that database. The limits of performance of the current disproportionality statistics are similar and new methods, involving substantially different approaches, may be required to gain appreciable improvements using spontaneous reporting data.

Recommendation	Rationale
Choice of a disproportionality statistic for signal detection should be primarily based on ease of implementation, interpretation and optimisation of resources	Several disproportionality statistics are currently used in data mining spontaneous report databases. All these can achieve similar overall performance by choice of appropriate signal detection algorithm. Thus, choice should be based on criteria other than signal detection performance. Factors that might be considered include the computing requirements to run the system, the ease of maintaining and adapting the system and whether the operation of the system can be easily communicated to non-statisticians [35]

Recommendation	Rationale	Recommendation	Rationale
Consideration should be given to the choice of signal detection algorithm used with disproportionality statistics because these can have important effects on quantitative signal detection performance	In contrast to the choice of disproportionality statistic, the choice of signal detection algorithm to define a SDR can provide very different levels of quantitative signal detection performance in terms of sensitivity, precision and time to signal. Hence, these criteria must be carefully selected on the basis of empirical evidence [35]	Consideration should be given to carrying out comparisons of quantitative signal detection methods across spontaneous report databases matching at the drug-event combination level rather than averaging over all drug-event combinations	It is possible that some ADRs may be more easily found in some databases. This was not investigated in PROTECT
For moderate to large spontaneous report databases, the relative performance of a quantitative signal detection algorithm in one database can be predicted from research in other databases	In the PROTECT study, signal detection algorithms with good signaling properties (in terms of sensitivity and positive predictive value) compared to other signal detection algorithms in one spontaneous report database also had relatively good signaling properties in other spontaneous report databases. The databases were both regulatory and company based and ranged in size from about 500,000 to 5,000,000 reports. Hence, relative performance in moderately large databases can be reliably inferred from evaluations in other settings [35]	It would be useful to conduct research to establish empirically the best method for quantitative signal detection in combination products	Combination products and single substances are often treated as unrelated in signal detection systems; a question remains whether combining data from these products will provide more or less accurate detection of signals
Absolute performance of the selected quantitative signal detection algorithm must be validated in the target spontaneous report database	Although the relative performance of signal detection algorithms is similar in different spontaneous report databases, the absolute performance characteristics may vary substantially. Hence, it is advisable to test the chosen disproportionality statistic with a range of signal detection algorithms within the target database [35]	Consideration should be given to establishing a framework for selecting the best quantitative signal detection algorithm to suit the organisational goals and resource available within a pharmacovigilance group	Our research has shown a predictable trade-off between sensitivity and precision as far as purely quantitative signal detection algorithms are concerned. However, the means of striking the correct balance between sensitivity and the concomitant burden of false positives for a given organisation requires careful consideration
Consideration should be given to the effect of reduced positive predictive value with time on the market	There appears to be a reduction in precision with time and hence it may be more productive to put additional effort into the evaluation of signals from newer products. This finding has been validated excluding ADRs identified prior to authorisation from the reference database but further work is ongoing to characterise this effect [35]		

2.5 Use of Subgrouping and Stratification in Disproportionality Analysis

Spontaneous report databases cover a range of products aimed at diverse medical conditions and used across a broad range of patient populations. This diversity is important as, for example, vaccines are given to healthy subjects, often children who are likely to have fewer underlying medical conditions and consequently different reported background adverse events than the main population of patients that use other medicines. Many quantitative signal detection algorithms disregard this diversity and give equal weight to information from all products and all patients when computing the expected number of reports for a particular drug-event pair, which may result in signals either being masked or false associations being flagged as potential signals. Stratification and subgroup analyses are generally used in epidemiology to reduce confounding and highlight effect modification. Both of these approaches may also have advantages in quantitative signal detection.

Other published studies have suggested some benefits of stratified and subgroup analyses but often the analyses included only a few key covariates or study products, and were conducted in single databases [36–44]. It is not clear how generalisable these results are to other spontaneous report datasets of different sizes and characteristics. Additionally, to our knowledge a head-to-head comparison of stratified and subgroup analyses against a reference standard has not been conducted.

The study of Seabroke et al. [45] investigated the impact of stratification and subgrouping in signal detection algorithms in spontaneous report databases of different sizes and characteristics using a range of key covariates (age, gender, calendar time period, country of origin, vaccines/non-vaccines, event seriousness, reporter type and report source). Signal detection performance was measured against a reference standard. Disproportionality analyses were conducted using either stratified or subgroup approaches and compared with an unstratified crude analysis. Stratified and subgroup analyses calculated disproportionality statistics within each individual stratum separately. For the stratified analyses, these were combined into a single value using a Mantel–Haenszel approach, whereas for subgroup analyses, a positive signal was counted if any of the individual strata met the signal criteria. The results were presented as sensitivity and precision (positive predictive value) for each approach calculated using a reference set of ADRs compiled from the SPC and company core data sheets as a proxy for true positives. Additional analyses included investigating the benefit of combined subgroup/stratified variables and also investigating the impact of a permutation analysis that used randomly split strata of equal size to a real variable of interest and compared the results with those for the real variable.

Whilst the spontaneous report databases employed in this study included large international, national and industry datasets, the results may not be generalisable to all spontaneous report databases, particularly those with a small volume of reports. The results from this study showed that subgroup analyses consistently performed better than stratified analyses in all databases. Subgroup analyses were also shown to provide clear benefits over crude analyses for some databases whilst stratified analyses were not found to increase either sensitivity or precision beyond that associated with analytical artefacts of the stratified analysis.

Recommendation	Rationale
Subgroup analyses may be beneficial in routine first-pass signal detection and should be considered. Stratified/adjusted analyses are unlikely to provide added value	In spontaneous report databases with over 0.5 million reports with broad diversity of products, subgroup analyses tended to perform better than stratified/adjusted analyses in all spontaneous report databases. Stratified/adjusted analyses were not found to increase either sensitivity or precision beyond random variation [45]
Subgroup analyses can be considered beneficial in large international spontaneous report databases with over 2 million reports. Smaller datasets especially those with reports from only one country may need to consider a likely tradeoff between increased precision with some loss of sensitivity if subgroup analysis was to replace a crude or adjusted analysis	Subgroup analyses within the larger international datasets consistently showed benefits in both precision and sensitivity over crude analyses for two disproportionality methods/thresholds with differing performance characteristics. For the smaller spontaneous report databases, a gain in precision tended to result in some loss of sensitivity particularly for the stricter disproportionality method/threshold and for the regulatory dataset with reports from only one country [45]
Subgrouping by seriousness of ADR or routinely excluding legal cases is unlikely to provide benefits in signal detection in terms of increased sensitivity or precision	Subgrouping by seriousness of the ADR defined using the IME list[a] had little effect on sensitivity or precision in any spontaneous report database. An analysis excluding cases submitted by lawyers also had little effect in all spontaneous report databases apart from the largest international database, which showed an increase in sensitivity and precision when legal cases are excluded [45]
Subgrouping by gender, reporter type and 5-yearly time points may provide modest improvement in precision in all, and sensitivity in some, spontaneous report databases	Subgrouping by gender, reporter type and 5-yearly time points showed a modest improvement in precision for all spontaneous report databases and improved sensitivity for larger and international databases. Implementation of these subgroup analyses into routine signal detection may provide some benefit [45]

Recommendation	Rationale
Subgrouping by age, country or continent of origin, or a combination of these variables, may confer improved precision in all and enhanced sensitivity in some spontaneous report databases	Subgrouping by age, country of origin, continent of origin and a combination of these variables showed the highest improvement of precision in all spontaneous report databases and sensitivity in the larger databases. Implementation of these subgroup analyses may be beneficial in optimising quantitative signal detection [45]
Subgrouping by vaccines/non-vaccines should not be implemented without careful consideration of the desired effect	Subgrouping by vaccines/non-vaccines resulted in a decrease in both precision and sensitivity in all spontaneous report databases. This was almost exclusively driven by the vaccines subgroup. These effects were owing to the suppression of listed vaccine ADRs as a result of comparing vaccines to each other. This may be desirable for certain reactions e.g. injection-site reactions but undesirable for other more serious reactions e.g. Guillian–Barre syndrome [45]
Where subgrouping by variables with considerable missing data (e.g. age, gender) is undertaken, consideration should be given to including a stratum for unknown rather than excluding these cases	Including missing data in the subgroup analyses for age and gender increased sensitivity in all spontaneous report databases but tended to also decrease precision. In spontaneous report databases with higher levels of missing data (20+ %) the increase in sensitivity was greater than the decrease in precision [45]
Subgrouping with a threshold based on number of reports may benefit from basing the threshold on the entire drug-event combination rather than within each individual stratum	Results for subgroup analyses that used an overall threshold of n applied to the whole drug event combination showed large increases in sensitivity but with loss of precision. Further validation would be needed within each organisation to ascertain whether this approach is sustainable in respect of resources available to evaluate an increased number of false positives [45]

Recommendation	Rationale
Future research should evaluate the use of subgroup analysis in parallel with crude and/or adjusted analysis	Results for subgroup analyses that used an overall threshold of n applied to the whole drug event combination showed large increases in sensitivity but with loss of precision. Further validation would be needed within each organisation to ascertain whether this approach is sustainable in respect of resources available to evaluate an increased number of false positives [39]

[a] The EMA Important Medical Event Terms (IME) list (https://eudravigilance.ema.europa.eu/human/textforIME.asp)

2.6 Influence of Masking on Disproportionality Analysis

Disproportionality analysis uses the spontaneous report database itself as the basis for computing an expected number of reports on a particular drug and adverse event. This is based on the assumption that true causal relationships between drugs and events do not influence the overall reporting rates and that the degree of under- (or over-) reporting for a given event is approximately the same for all drugs. In practice, this assumption may be invalid for some drugs and events in a spontaneous reporting database. For example, attention to a real or perceived safety issue in the medical community or in public media may increase the reporting rate for that drug-event pair to such an extent that the overall reporting of that event is affected and potentially masks alerts from other drugs [46–51].

The two studies that made up this PROTECT work package [49, 50] looked at the impact of masking on disproportionality analysis. An algorithm was used to generate a masking ratio whereby the masking effect of one drug on another can be estimated for a specific event and suggests a simplified version valid under certain conditions. A simulation study was performed that focussed specifically on comparing differences between the simplified and exact masking ratio [49]. The impact of different approaches to treating the reports was also explored, including both the suspected masking drug and the suspected masked drug as concomitant medicines. In a follow-on study, the extent and impact of masking was studied in two spontaneous

reporting data sets: EudraVigilance and Pfizer's spontaneous report database [50]. The latter evaluated the impact of removing from the analysis for each ADR the drug with the highest masking ratio. It was found that the drugs inducing the highest masking effects tended to be those that are known to cause the ADR in question. Under the conditions of the study (assuming that each ADR is masked by exactly one drug), it was rare that the unmasking analysis affected whether a drug-event pair was considered to be disproportionally reported or not. However, it is important to note that the drug-event pairs affected in this way primarily involved rarely reported ADRs.

Recommendation	Rationale
Quantification of the masking effect of drugs on adverse reactions or adverse reactions on drugs could be used as a diagnostic tool of the extent of masking at two levels: For determining, based on the general characteristics of a spontaneous report database, if the application of an unmasking algorithm would be worthwhile At the product-event pair level, if a specific concern is raised about a potential masking effect driven by another product or another group of products	Results indicate that many drugs and adverse reactions are not affected by masking. Avoid complicating the analysis of data by adding an unmasking procedure when masking is not an issue. Formulas for assessing the effect of masking can be found in papers by Maignen et al. [49, 50]. As the characteristics of the spontaneous report database change over time, it is still interesting to monitor the extent of masking periodically. During signal evaluation, some evaluation of the masking effect could be performed at the level of the product-event pair [49, 50]
If the masking effect of drugs on adverse reactions or adverse reactions on drugs is substantial, applying an unmasking algorithm should be considered	Reducing the effect of masking can increase the sensitivity of quantitative signal detection and, in principle, result in earlier identification of new drug-event associations [49, 50]
If false negatives are a major concern, unmasking of drugs and/or adverse reactions can be used in parallel with standard disproportionality analysis to improve sensitivity and timeliness but this benefit must be balanced against the cost in increased evaluation of false positives	If unmasking and standard disproportionality analyses are used in parallel, sensitivity will be equal to or higher than that of standard disproportionality analysis alone, but parallel analyses of the data also increase the false-positive rate, from spurious associations [49, 50]
Future research should explore the effectiveness of unmasking in terms of true/false positives revealed by an algorithm	In the absence of public health evidence from prospective studies on the benefits of removing the masking (or situations in which unmasking could be beneficial), the use of a particular algorithm should be directed by the rate of true signals/false positives revealed by the removal of the unmasking effect [49, 50]

continued

Recommendation	Rationale
Future research should compare disproportionality-based approaches for unmasking to other statistical approaches (e.g. logistic regression models) that could also be used to account for masking effects	This was outside of the scope of the PROTECT studies and there appears to be no published research on this topic
The use of simple unmasking algorithms as a means of reducing computation complexity and improving transparency should be explored in a future study	Results indicate that the performance of the simplified methods is comparable to that of more complex methods while the computational complexity is reduced and transparency improved, but further research is needed to fully explore this on datasets with different properties [49, 50]

2.7 Drug–Drug Interaction Detection

Adverse drug–drug interactions (DDIs) harm large numbers of patients every year. Not all DDIs are known when new medicines are made available to the general public, but individual case reports may enable post-marketing detection. Earlier studies have indicated that statistical measures for DDIs that use additive baseline models perform better than those that use multiplicative baseline models [52, 53] but no broad evaluation has been reported in the literature. A study was conducted to compare the sensitivity and specificity of different statistical measures for ADR detection against established and emerging adverse DDIs, respectively.

Analyses were performed in VigiBase where four statistical measures for DDI detection were evaluated: one based on regression with a multiplicative baseline model [53], one based on regression with an additive baseline model [53], one shrinkage disproportionality statistic with a multiplicative baseline model [52] and one shrinkage disproportionality statistic with an additive baseline model [52]. The reference set for known interactions consisted of 74 established DDIs and 29 pairs of drugs with an ADR for which there was no empirical support for a DDI. The reference set for emerging DDIs included 324 adverse drug interactions added to Stockley's Interaction Alerts between 2007 and 2009, and 324×20 combinations with two drugs that were not listed together as known to interact in the same reference, in 2009. The majority of the ADRs were investigated at the Meddra PT level.

The study was limited to statistical interaction measures, whereas recent research has suggested that predictive models accounting for multiple aspects of strength of evidence may perform even better [54]. The algebra of the

statistical interaction measures is such that additive models impart superior sensitivity regarding the detection of DDIs compared with multiplicative measures, at any given value for the threshold. However, the additive models used in this study also demonstrated better specificity compared with the multiplicative models. Thus, statistical interaction measures with additive baseline models outperformed those with multiplicative baseline models for both established and emerging adverse DDIs. However, given the small number of cases attributed to the interacting drugs together, both models gave low sensitivity (<20 %) for emerging adverse DDIs, at conventional signalling criteria.

Recommendation	Rationale
Statistical interaction measures with additive baseline models should be preferred over those with multiplicative baseline models for detecting signals of DDIs in spontaneous report databases	Statistical interaction measures with additive baseline models provided better sensitivity and equal or better specificity for both established and emerging DDIs [55]
Future research should explore how statistical interaction measures with additive baseline models can best be incorporated in broader predictive models of adverse drug interactions, and in routine signal detection	This was out of scope for the PROTECT study, but recent research has found that predictive models accounting for multiple aspects of strength of evidence perform better than statistical measure of interaction alone [54]

2.8 Duplicate Detection

Individual case reports of suspected harm from medicines are fundamental for signal detection in post-marketing surveillance. Their effective analysis requires reliable data and one challenge is report duplication. These occur when multiple unlinked records are recorded describing the same suspected ADR in a particular patient. Report duplication is known to occur for a diverse range of reasons, including reporting of an ADR from multiple primary sources, requirements for marketing authorisation holder reporting of literature cases (and subsequent retransmission), and technical or administrative issues [56]. De-duplication of data is also widely understood to be a time-consuming process; however, little research had been undertaken prior to PROTECT on the efficiency of different de-duplication methodologies. Duplicate ICSRs are known to distort statistical screening [56] both increasing the numbers of false-positive and false-negative signals. The net effect of duplicates (and de-duplication techniques) is unknown, and may not be consistent across all products in the database;

however, where there is a significant impact there is potential for the cases to mislead clinical assessment.

Many organisations rely on rule-based duplicate detection methods, which rely on exact matching of a number of elements within individual case safety reports and displaying the putative duplicate cases for review. Rule-based approaches can vary significantly in their complexity and based on the fields used, and may consider patient, reporter, drug or reaction details, or a combination of them. The study compared methods [57] used in the UK, Denmark and Spain. The UMC had previously published a probabilistic record matching algorithm that indicates the likelihood of cases being duplicates (vigiMatch [58]) as an alternative to rule-based approaches. The PROTECT study attempted to quantify the benefits (or disbenefits) of the different approaches used by PROTECT consortium members.

A first phase of the study aimed to evaluate probabilistic record matching for duplicate detection compared with rule-based approaches. Studies considered positive predictive value and numbers of false positives of different approaches, and in addition attempted to characterise the main causes of duplication. Initial research was undertaken using the WHO Global Individual Case Safety Reports Database, VigiBase, for reports submitted between 2000 and 2010. Suspected duplicates for the UK, Denmark, and Spain were reviewed and classified by the respective national centre. This included evaluation to determine whether confirmed duplicates had already been identified by in-house rule-based screening. A second phase of the study directly compared results from the MHRA's rule based approach to the UMC's probabilistic record matching approach using data received in the MHRA's Sentinel database during 2013.

Probabilistic record matching performed positively when compared with rule-based approaches. Specifically, probabilistic record matching demonstrated a high predictive value above that of rule-based methods and is expected to improve efficiency and accuracy of duplicate management. The study showed very few false positives suggesting it may be possible to increase sensitivity while ensuring false-positive rates are kept at a reasonable level. The study [57] highlighted that case management system changes or upgrades can occasionally result in large numbers of duplicates, either in the source system, or those external to the organisation. It was also noted that proliferation of duplicated cases within databases occurred as a result of rapid submission and re-transmission of cases to multiple stakeholders made possible by electronic systems. This emphasises the need for swift and robust duplicate detection procedures at each organisation. Data privacy requirements were noted as a barrier to the most effective

duplicate detection and it was considered that approaches to deidentifying data (for example, scrambling dates and patient initials) in a way that permits duplicate detection should be pursued to allow for effective duplicate detection in databases that pool reports from different sources.

Although beyond the scope of the PROTECT study, further evaluation should be undertaken to understand the feasibility and impact of automatic exclusion of potential duplicates from quantitative signal detection algorithms. Such an approach may help ensure independence of reports, which is a fundamental assumption underlying the computation of confidence intervals for all disproportionality statistics; however, it equally may remove clusters of reports from the same reporter that may be important for patient safety. A better understanding of the reasons for related cases that are not considered duplicates, and their scientific implications for signal detection will help determine if this is a viable approach.

Recommendation	Rationale
Probabilistic record matching should be considered as an alternative to rule-based methods for duplicate detection in pharmacovigilance	Probabilistic record matching demonstrated a high predictive value above that of rule-based methods in our study, and is expected to improve efficiency and accuracy of duplicate management [57]
Care should be taken to avoid case duplication during system changes/upgrades, considering both internal aspects and case transmission to external organisations	Our study showed that such changes on occasion resulted in very large numbers of duplicates [57]
Rapid electronic re-transmission of spontaneous adverse drug reaction reports between databases can increase the number of duplicates to the extent that disproportionality statistics are are significantly affected, emphasising the need for swift and robust duplicate detection and management processes in databases that employ electronic data exchange	There are a large number of duplicates in spontaneous reporting databases, which are shown to affect quantitative signal detection scores. Rapid transmission of cases by electronic systems exacerbates this issue, meaning that accurate (and ideally, non-burdensome) duplicate detection processes are required to mitigate this unwanted impact on disproportionality statistics [57]
Further work should be undertaken to explore lowering the threshold for the tested probabilistic record matching method and methods in general to evaluate the balance of false positives and negatives	Our study showed very few false positives, so it should be possible to increase sensitivity while ensuring false positive rates are kept at a reasonable level [57]

Recommendation	Rationale
Further evaluation should be done to understand the impact of automatic exclusion of potential duplicates from quantitative signal detection algorithms	This was beyond the scope of the PROTECT study. If this approach proved successful manual duplicate detection activities could be eliminated resulting in time/resource savings [57]
Approaches to deidentifiying data (for example, scrambling dates and patient initials) in a way that permits duplicate detection should be pursued to allow for effective duplicate detection in databases that pool reports from different sources	This will reduce the negative impact of data privacy laws, for duplicate detection in international databases [57]

2.9 Relationship Between Disproportionality Measures (i.e. PRR) and Risk Estimates

Spontaneous reports are only submitted when a patient exposed to a medicine experiences a suspected ADR and the reporter decides to report the case. It is difficult to determine the extent to which the numbers of these reports may help to reflect a measure of association of adverse outcomes in the drug-exposed population. Thus, the quantitative data in spontaneous reporting systems, while being useful in detecting new signals of drug-event associations, are not easily interpretable in terms of clinical impact. Nevertheless, quantitative signal detection systems consider threshold levels for disproportionality measures and therefore higher values of disproportionality statistics are one of the factors that influence the decision to investigate particular drug-event combinations. Hence, it makes sense to ask whether there is a direct relationship between the magnitude of disproportionality statistics and the magnitude of the association between a product and an adverse effect from pharmacoepidemiological studies.

A study was conducted to determine the proportional reporting ratios (PRRs) for a set of known ADRs and compare them with estimates of association from formal epidemiological studies [59]. A set of 15 confirmed ADRs were selected from the initially identified dataset of pharmacovigilance driven European Union regulatory actions and for which relative risk estimates from formal studies were available and were considered to provide

well-established evidence supporting the respective regulatory actions. Prior to any calculation of PRR, the studies were collated and a best estimate of the risk ratio selected on the basis of pre-specified rules. When available, this estimate was that determined during the European Union regulatory assessment. At the same time, an estimate was made of the date at which each ADR was first publicly recognised.

Only after the risk ratio was decided upon was the PRR calculated by reconstruction of the spontaneous reporting data system at the predetermined date of first recognition in the medical community of the ADR. The primary analysis used the EudraVigilance spontaneous report database and an additional analysis was carried out in FEDRA, the Spanish national spontaneous report database. Case definitions for the ADR of interest were developed for PRR calculations for each drug-event pair, aiming to reproduce the case definition as used in the epidemiological studies providing relative risks. An initial dataset of 78 drug-event pairs was obtained following the initial inclusion criteria. Following the exclusion criteria, 15 drug-event pairs were finally selected. The pairs include 13 different ADRs and 14 different drugs/classes. Four of the ADRs represented class effects. Four topics were related to outcomes of safety referrals concluding in a CHMP Opinion and 11 were recommendations from the Pharmacovigilance Working Party to CHMP or to National Competent Authorities. Eleven of 15 drugs were non-centrally authorised medicines and four were centrally authorised in the EU.

An orthogonal regression model showed a significant association between relative risks and PRRs. This suggests that, in some cases, there is a relationship between the PRR calculated immediately prior to first awareness of the safety topic and estimates of relative risk taken from published epidemiological studies, where the signal turned out to be confirmed. It is noted that no validation of the model has been performed and that despite the good correlation shown between RRs and PRRs in this exercise, it is emphasised that PRR cannot replace RR. Thus, calculation of PRRs from spontaneous reporting databases should not replace nor delay the performance of formal epidemiological studies but could however be an indicator of the likely clinical importance of the adverse reaction, should the signal be confirmed subsequently [60].

Recommendation	Rationale
It may be possible to use the PRR at the early phase of the analysis of a new safety signal as an indicator of the likely strength of the association, should the signal be confirmed	The PRR observed before general awareness of an ADR shows a good correlation with the strength of the association in terms of relative risk or odds ratios later established by controlled studies. However, the PRR is not a direct estimator of the risk ratio and should be considered only in the absence of any more reliable evidence. The caveat 'should the signal turn out to be confirmed' must be observed. The study analysis does not compare the distribution of PRR values for 'true' and 'false' signals of disproportionate reporting and no inference can be made about whether the initial magnitude of PRR gives information about the nature of the association (causal or otherwise) [59]
Following the initial detection of a signal of a specific drug-event association, PRR values based on clinical definitions of the adverse event may serve to provide an estimate of the likely size of clinical effect and be included among the criteria for initial prioritisation of its assessment	This study shows that, at least in this selected set of study cases, the underlying relative risk seemed to influence both the direction and magnitude of the PRR calculated with a similar case definition of the adverse event. Because the study sample comprises drug-event associations confirmed following assessments of diverse data sources and signal detection systems, the results may be applicable to PRRs calculated following both quantitative and traditional signal detection approaches [59, 60]
Consideration should be given to repeating these analyses in other ADR datasets to see whether they can be replicated and, if they can be, to establish the relevant scale factor	PRR values generated in different ADR datasets are unlikely to be the same. Other IMI PROTECT research has focussed on the performance of disproportionally statistics and of different signalling algorithms in different ADR datasets. However, to date no attention has been paid to describe and explain differences in the calculated PRR values in different datasets

Recommendation	Rationale
Consideration should be given to further exploring whether PRRs adjusted by subgroup variables improves the correlation with measures of association from studies	The findings from an IMI PROTECT study on sub-grouping and stratification [45] suggest that subgrouping may be a useful strategy to try to improve the correlation between the PRR and the estimates of risk from studies
Consideration should be given to exploring whether PRRs calculated for single MedDRA® PTs as is in EudraVigilance monitoring behave in the same way as the clinically defined case definitions in terms of correlation with measures of association from studies	The medical concepts used in the studies to derive the estimates of relative risk often described broader medical concepts than the MedDRA® PT level used in EudraVigilance for the PRR screening analysis (see also Recommendations in relation to Timeliness of Quantitative Signal Detection using MedDRA® Terms and Groupings)

2.10 Signal Detection in Longitudinal Observational Data

Post-marketing surveillance aims to identify and characterise risks of medicines. At present, signal detection is predominantly based on individual case reports, but the use of electronic health records and insurance claims to detect ADRs is an area of active research [61–64]. While some work has focussed on traditional sequential approaches and looking to use them for signal detection, new methods have also been adapted or proposed [65–68]. Other publications have highlighted the challenges and limitations of using longitudinal observational data for signal detection [69, 70].

The objectives of the three studies that made up this PROTECT work package were (1) to better characterise the opportunities and challenges for prospective signal detection in longitudinal electronic health data, including the development and evaluation of a process for the structured assessment of potential safety signals from electronic health records; (2) a comparison between exploratory and confirmatory analyses of longitudinal electronic health data and; (3) a performance evaluation of quantitative signal detection in longitudinal electronic health data and individual case reports, respectively, for emerging safety signals. All analyses were performed in The Health Improvement Network (THIN[4]) database of longitudinal electronic health records from general practices in the UK.

---

[4] THIN is an electronic medical record data resource including over 12 million individual patients from the UK, with over 3.8 million currently active patients. The electronic medical records are collected from general practices in primary care (http://www.ncbi.nlm.nih.gov/pubmed/22828580).

.

The comparison between exploratory and confirmatory analysis was based on 13 published studies of possible ADRs in THIN [71]. The selected studies listed a total of 56 drug-event pairs that were included in the analysis. For each pair, the analysis results of the study (positive or negative) were compared with those that would result from analysis with standard design choices for a calibrated self-controlled cohort analysis in THIN. Observed differences were closely scrutinised to identify areas of possible improvement for the standard implementation of the design, for exploratory analysis. In our comparison to published epidemiological studies, a common discrepancy was that the epidemiological study performed analysis for all drugs in a class together and/or for a number of related medical events together, which improves power. However, our detailed review often found substantial and important differences among different drugs in the same class or among different medical events in the same category. Clearly, more research is needed to minimise terminological obstacles for exploratory analysis of longitudinal data, as well as across the different data streams [70]. None of the false positives were considered to represent a chance association: all were considered to be the result of systematic variability.

For prospective identification of potential signals in electronic health records [72], a questionnaire for structured assessment was designed and iteratively refined through pilot testing. It covered aspects such as the nature of the temporal pattern, the presence of co-medications associated with the medical event, the likelihood of confounding by underlying disease, and other alternative explanations for observed temporal associations. In real-world use, its purpose would be to provide analysis and decision support for potential signals identified through prospective and open-ended screening of longitudinal electronic health data. For the purpose of the study, drug-event pairs temporally associated according to a calibrated self-controlled cohort analysis in THIN were randomly selected for review. Six assessors trained in pharmacovigilance and/or epidemiology participated in the main study and each evaluated up to 20 temporally associated medical events per drug, for seven randomly selected drugs [72]. Our analysis highlighted a number of potential safety signals in electronic health records that merit further review. These range from life threatening to those that are less serious, but important for patients and for adherence. However, three out of four temporal associations identified in the initial screen could be dismissed from further evaluation after the initial review. In other words: without a review, the majority of the highlighted associations would have been false positives. A minority of the dismissed associations were considered to be owing to random variability; most were the result of biases and other systematic effects [72].

The performance evaluation of quantitative signal detection in longitudinal electronic health data and spontaneous reports was based on a reference set of 264 historical safety signals derived from a previously published study [11], and 5280 negative controls. The literature is very sparse on such comparisons. One example is from Trifirò et al. [73], who as part of the EU-ADR (European Commission sponsored project 'Exploring and Understanding Adverse Drug Reactions by integrative mining of clinical records and biomedical knowledge'), looked at the number of drug-event combinations that were highlighted as disproportional in each of the data sources. Analyses were performed in THIN and VigiBase, each backdated to the end of 2004, to match approximately the time of the historical signals. The analysis in VigiBase was repeated with a restriction to reports originating only from the UK. The retrospective evaluation against historical safety signals for European centrally authorised products showed that none of them could be detected in THIN with the method we used, prior to the initial signal at the EMA. In many cases, this was because of the drug not being available on the UK market at the time, or the drug or medical event not being reliably captured in primary care. In contrast, some of the positive controls could be detected in VigiBase, even when we restricted the analysis to individual case reports originating from the UK.

Recommendation	Rationale
Longitudinal observational data should be further explored as a complement to signal detection using individual case reports but cannot currently replace individual case reports for this purpose	Individual case reports of suspected harm from medicines have a proven value for safety signal detection. However, they are not optimal for detecting increased rates of multifactorial adverse drug reactions with high background incidence. Longitudinal observational data provide the basis for epidemiological evaluation of such associations and should in principle enable their initial identification. However, we lack evidence to suggest that signal detection in longitudinal observational data can match the performance of signal detection in individual case reports for all drugs and medical events. In our evaluation of historical safety signals from the EMA, none of the positive controls could be detected in the THIN database at an early stage, whereas this was possible in VigiBase for some of the signals, even when we considered only the subset of the UK individual case reports [72]

Recommendation	Rationale
Safety signal detection in longitudinal observational data should include clinical, pharmacological and epidemiological review of identified temporal associations	Clinical review of statistical signals is fundamental in evaluating signals arising from spontaneous report databases. In our study of structured assessment for prospective identification on safety signals in electronic health records, three out of four temporal associations identified in the initial screen could be dismissed from further evaluation after initial review. Without review, the majority of the highlighted associations would have been false positives [72]
To the extent possible, temporal associations detected in longitudinal observational data should be further explored with statistical graphical methods	In our prospective identification study, in-depth review of the chronograph temporal patterns proved a valuable component of the expert review. Univariate measures of temporal association may over-simplify or obscure the underlying patterns in such rich, complex and often long records [72]
Safety signal detection in longitudinal observational data should account for limitations of the underlying data and take measures to ensure appropriate interpretation. In selecting the data set for analysis, one should account for both its size and scope (which drugs and diagnoses it captures) and for the fact that effective review of identified temporal associations requires expert knowledge of the underlying data, which is particularly relevant for large heterogeneous data sets	Our retrospective evaluation against historical safety signals for European centrally authorised products showed that none of them could be detected in THIN with the method we used, prior to the initial signal at the EMA. In many cases (to be further specified once we have the data), this was because of the drug not being available on the UK market at the time, or the drug or medical event not being reliably captured in primary care
Future research should explore the relative merits of performing safety signal detection in longitudinal observational data for groups of medicinal products and medical events, instead of or in parallel with that of individual products and events	In our comparison to published epidemiological studies, a common discrepancy was that they performed analysis for all drugs in a class together and/or for a number of related medical events together, which improves power. However, our detailed review often found substantial and important differences among different drugs in the same class or among different medical events in the same category.

2.11 Signal Detection in Clinical Study Data

In the pre-approval phase of a drug, systematically collected adverse event data from randomised clinical trials are the principal data source and therefore the cornerstone of safety analysis. Specifically, clinical trial data allow the estimation of exposure, which is one of the limitations of spontaneous reporting datasets [7]. In addition, randomisation itself is a powerful feature of randomised controlled trials that addresses the issue of confounding, which otherwise complicates attempts to identify imbalances in the incidence of adverse events between the treated and untreated subjects in data from non-randomised sources. Because of this, clinical trial data form a natural source for signal detection, especially in the early phase of a drug's lifecycle.

Two different types of analyses performed in company clinical trial databases are presented: in the first, the use of extreme value modelling for the prediction of potential drug toxicity was evaluated; in the second, different approaches for dealing with multiplicity issues for adverse event-based signal detection were compared.

It is often the case that potential drug toxicity is suggested by the occurrence of extreme values of clinical variables, rather than changes in the location of the distribution of these variables. For example, large values of serum creatinine and blood urea nitrogen are used in the diagnosis of acute renal failure, large values of the QT interval are suggestive of cardiotoxicity and large values of alanine aminotransferase suggest hepatotoxicity. Many other examples of abnormally large (or small) values of laboratory variables being indicative of safety issues can be found in the Common Terminology Criteria guidelines,[5] and, in general, any laboratory values that fall above the upper limit of normal or beneath the lower limit of normal may be indicative of a potential safety issue. Often, extreme values are poorly predicted by an analysis that focuses on the central features of the distribution. Therefore, analyses of means or medians tend to be uninformative. Extreme value modelling might therefore provide a novel approach to the prediction of drug toxicity early on during drug development.

A retrospective analysis of phase II data using extreme value modelling was conducted for ximelagatran [74], a compound that was denied marketing approval in the USA and was withdrawn from those markets in which it had been approved because of concerns over potential liver toxicity. The analysis showed that the phase II data were predictive of the phase III results and, had the methods been available at the time, such analyses would have provided valuable information relating to the decision to proceed with further development of the compound.

One important characteristic of signal detection in clinical trials as well as in other databases is a focus on the evaluation of a large number of endpoints. Analyses of adverse event data typically generates multiple risk measurement estimates associated with events across several body systems. Selection of outcomes for further evaluation by conventional hypothesis testing of between-group differences for each endpoint can be problematic. Ignoring the fact that the data determine the hypotheses that are tested, by using a process that involves a large number of comparisons across multiple analysis time points multiplicity becomes an issue. Acknowledging P-values as a useful flagging device, ICH E9 recommends statistical adjustments for multiplicity when applying hypothesis tests to a large number of safety variables in clinical trial data. However, probably owing to the concern of missing true safety signals (i.e. false-negative signals), multiplicity adjustment has also been described as counterproductive and again probably because of these and similar reservations, adjustments are often avoided, despite the fact that ignoring multiplicity may easily result in an excessive rate of false-positive signals. The EudraVigilance Expert Working Group notes that thresholds commonly used to detect signals in spontaneous data are a trade-off between two conflicting goals: "either generating too many false positive signals if the threshold is too low or missing true signals if this threshold is too high". Given this trade-off, it is important to identify and calibrate methods to strike a reasonable balance between these two parameters.

The objective of this study [75] was to investigate different approaches to address multiplicity for the use of signal detection methods to select ADR candidates in clinical trial data. The aim was to identify the best performing method that maximises the proportion of correct signals (i.e. the positive predictive value) as compared with the CDS as the gold standard. In addition, the use of different MedDRA® levels for signal detection reflects the importance of considering the MedDRA® hierarchy for signal generation (see above, 'Timeliness of Quantitative Signal Detection using MedDRA® Terms and Groupings'). In addition to basing the analyses on the PT as the smallest unit of analysis, the use of available company-specific grouping of PTs developed specifically for the purpose of labelling so-called medical labelling groupings was evaluated.

As expected, the ability for ADR detection was highly influenced by ADR frequency. In general, all model types that took multiplicity into consideration proved appropriate for the detection of signals. The Bayesian hierarchical model that can make use of the hierarchy, thereby borrowing strength, performed best among the quantitative

---

[5] Common Terminology Criteria for Adverse Events, US National Cancer Institute (http://ctep.cancer.gov/).

signal detection methods, especially with regard to time to signal. In our analysis, the use of medical labelling groupings did slightly improve the performance of quantitative signal detection algorithms.

In summary, taking multiplicity into consideration resulted in lower rates of false-positive signals. The hierarchical Bayes and double-FDR, which also account for the hierarchical structure of the underlying medical terminology dictionary, demonstrated the best performance in the investigated setting. Considering the good positive predictive value while providing similar sensitivity, these methods can provide an alternative to the often-used unadjusted analysis for identifying or flagging potential imbalances between treatment arms and as such could be used in the detection of ADR candidates. However, selection of the most appropriate methods must consider the size of the clinical trial database and computational requirements.

Recommendation	Rationale
If prior knowledge suggests data from a particular organ system should be monitored, consider extreme value modelling on data arising from each trial for the compound of interest. For example, if preclinical data suggested a potential liver issue, prepare to model ALT; if another compound in the class showed kidney signals, prepare to model creatinine	Extreme-value modelling has been demonstrated, in various examples, to provide useful predictions of drug toxicities from early-phase data. If it is possible to pre-specify the modelling and prediction exercise, the results have greater credibility than if they are data driven, and resources can be allocated up front to ensure the work is done to appropriate deadlines [74, 76]
Some analyses will be data driven, suggested by observed extremes in the data. These could also be subjected to extreme value modelling and the statistical evidence thus acquired interpreted in context	Not all potential safety issues are known in advance, so some analyses are necessarily data driven. It is inappropriate to consider such analysis illegitimate or to yield unreliable results provided they are interpreted in context. Statistical inference is only one part of the larger process of scientific inference [74]
Extreme value modelling can commence as early as phase I; however, in most cases, phase II data need to be available for reliable inferences to be made	Experience suggests that phase I data may be sufficient for extreme value modelling to identify toxicity, but that sometimes the sample sizes are too small. Modelling and prediction have the most value to add when the volume of available data is low, so such exercises should be commenced as soon as possible [74]

Recommendation	Rationale
Properly trained, Independent Data Monitoring Committees or Safety Review Boards are likely to benefit from extreme value modelling of unblinded data	When an IDMC exists, there are sometimes reasons for additional monitoring. It follows that applying proven methodology to emerging data will provide the best chance of identifying and characterising the safety issue as soon as possible [74]
When extreme value modelling does not find evidence of a safety signal in studies of short duration, extrapolation beyond observed durations of exposure is discouraged	It is reasonable to expect that some toxic effects of drugs will not manifest themselves until several weeks or months of exposure have occurred. If extreme values are not observed at relevant doses in short trials, proceed with caution, acknowledging that they could occur after longer durations of exposure
Multiplicity adjustment provides a useful tool to improve the positive predictive value in signal detection in clinical trial data. The use of multiplicity adjustment needs to be evaluated against the size of the available clinical trial database	The ability for ADR detection is highly influenced by ADR frequency in the source dataset. Thus, database size and event reporting frequency must be taken into consideration when the use of multiplicity adjustments for ADR candidate selection is considered [75]
The use of Bayesian Hierarchical Models can improve the efficiency of signal detection through borrowing of strength from other relevant events in the clinical trial dataset. This must be weighed against the more complex computational requirements of Bayesian modelling	Bayesian Hierarchical Models provided the best performance with regard to positive predictive value, specificity, sensitivity and negative predictive value, mainly owing to their ability to "borrow strength" across similar terms [75]
The use of more specific MedDRA® groupings can further improve signal detection in clinical trial data	The use of narrow-term groupings for analysis provided slightly better results for signal detection compared with the analysis based on MedDRA® PTs alone [75]

3 Discussion

These recommendations are presented to highlight the outcomes of the research conducted under the auspices of IMI PROTECT Work Package 3 and in such a way that pharmacovigilance professionals, particularly those with an interest in research and methods development can readily adopt appropriate potential improvements in their quantitative signal detection practices. They should not be considered a comprehensive treatise on the subject of quantitative signal detection but should be considered by

readers within the context of the entire body of signal detection research and guidance documents that exist on good signal detection practice [7, 77] outside of IMI PROTECT. For example, PROTECT's focus on quantitative signal detection methods means that recommendations relating to the relative merits of quantitative vs. qualitative signal detection methods cannot be made and pharmacovigilance organisations are likely to have to continue with both approaches to maintain an optimal signal detection capability. Although a broad cross-section of databases [78] and data sources were employed in the work package, the generalisability of recommendations to other databases and data sources should be considered carefully before implementing them, particularly in databases of adverse event reports smaller than those used in this project.

An important strength of PROTECT's signal detection research was the execution of standardised analysis protocols across multiple spontaneous reporting datasets [35, 45]. Several of these studies compared the databases of pharmaceutical companies, national regulatory authorities, and international organisations, such as the EMAgency and the WHO.

PROTECT found no overall benefit in conducting signal detection using MedDRA® HLTs or SMQs compared with using individual PTs [9]. Some relatively minor gain in time to signalling was seen when closely related (in a clinical sense) ADR terms were grouped together, an area for potential future research. This is compatible with an earlier study of the US FDA Adverse Event Reporting System, which found that analysis at the level of HLTs or SMQs decreased sensitivity but increased specificity. Its reference dataset were drug-event combinations with some degree of support in the literature that were not on the original drug label, but it did not consider timeliness [8]. PROTECT showed that knowledge engineering techniques can be used to derive novel groupings of adverse event terms dynamically based on the relevant semantic dimensions for a specific topic of interest, although no net improvement in signal detection performance was seen in these studies [16, 17]. It is interesting to speculate whether it may be possible to go beyond this restricted example and generate alternatives to the standard groupings available in existing adverse event terminologies, such as MedDRA® HLTs or SMQs.

One of the tangible outcomes of PROTECT work package 3 is the structured database of MedDRA® coded ADR information taken from section 4.8 of the SPC for all products authorised in Europe through the centrally authorised procedure [25]. This database is publically available on the EMA website and is maintained. It reduces the need for manual inspection of SPCs when the focus of safety monitoring is detection of new risks. Its use has

already been tested and implemented into signal detection processes at the EMA and the UMC.

In a broad study across spontaneous report databases, PROTECT showed that the choice of signal detection algorithm (e.g. threshold on the number of reports, the threshold on the disproportionality statistic, and/or statistical significance) was much more important than the choice of disproportionality statistic itself [35]. Performance of any single algorithm might be very different between one spontaneous report database and another but the relative performance of two algorithms was generally similar across different databases. In a related study across almost the same range of databases, the use of stratification and subgroup analysis in disproportionality analysis was explored [45]. An earlier study by Caster et al. [37] showed that the use of subgroup analyses and stratification both out-performed crude disproportionality analysis although the relative contributions of each approach were not determined. In the PROTECT study, it was shown that subgroup analyses tended to be more beneficial over stratified analyses across all datasets studied. Subgroup analyses also provided clear benefits over crude analyses in some datasets whereas stratified analyses did not increase sensitivity or precision beyond random variation. This unexpected finding that has not been reported elsewhere has important implications because a number of organisations routinely use stratification in their analyses, but very few routinely employ subgrouping. Other previous studies [36, 38] have observed modest improvements for stratified analyses consistent with the results from the stratified analyses in the PROTECT study and therefore these findings from the other studies may also be artefacts from the stratification process rather than a true effect.

PROTECT explored the impact of masking on statistical signal detection in spontaneous report databases [49, 50]. Under the conditions of the study, it was rare for masking to affect whether a drug-event pair was considered to be disproportionally reported or not; however, the drug-event pairs that were affected in this way primarily involved rarely reported ADRs. Furthermore, the study only considered removal of single masking drugs from the calculation and in some cases, multiple masking drugs may be present [51].

For the detection of adverse DDIs, PROTECT showed that statistical interaction measures with additive baseline models outperform those with multiplicative baseline models [55], typically available in standard statistical software. This finding was true for both established and emerging DDIs. Notably, for emerging adverse DDIs, the statistical interaction measures with multiplicative baseline models that might be the easiest to implement performed worse than would expected by chance.

Many organisations rely on rule-based methods for the detection of duplicate individual case safety reports; however, PROTECT showed that probabilistic record matching performed better than rule-based screening [57] and should be considered as a viable alternative. Specifically, probabilistic record matching demonstrated a high predictive value above that of rule-based screening, and is expected to improve efficiency and accuracy of duplicate management. This has important resource and quality implications for organisations where large volumes of case reports are exchanged on a routine basis.

A comparison between estimates of association from formal epidemiological studies and proportional reporting ratios in spontaneous reporting data for a set of known ADRs found a correlation, at a point in time before the ADR was first publicly recognised [59]. This study suggests that it may be possible to use the proportional reporting ratio at the early phase of the analysis of a new safety signal as an indicator of the likely strength of the association, should the signal be confirmed. Acknowledged limitations exist in the current evidence base for this association and have been discussed in an earlier publication [60].

At present, signal detection is predominantly based on spontaneous reports, but the use of longitudinal electronic health data in pharmacovigilance is an area of active research. Most studies to date have focussed on the statistical evaluation of well-established ADRs [79, 80] but not sought to define processes for effective identification of emerging safety signals in longitudinal health data. Similarly, comparisons between individual case reports and longitudinal health data for signal detection have focussed on established and not emerging ADRs. Broad studies such as those performed by observational medical outcomes partnership and EU-ADR have explored the merits of different epidemiological designs when applied automatically across broad ranges of drugs and outcomes [80–83] but such studies have been primarily retrospective in nature, and there is also a lack of studies to determine the relative merits of exploratory screening vs. customised confirmatory analyses of longitudinal health data. Some research has proposed signalling showing outputs requiring highlighting in both spontaneous reports and observational data [84] but without attempting to assess the relative value of the two data sources.

PROTECT performed research on statistical signal detection in the THIN database of longitudinal electronic health records from general practices in the UK. A process for structured clinical and epidemiological assessment of temporally associated prescriptions and events in electronic health records was developed and evaluated. It showed that important potential safety signals can be identified in these data, whereas clinical and epidemiological review of highlighted statistical associations is crucial to attain an acceptable false-positive rate [72]. Conversely, a retrospective evaluation did not detect any of about 500 historical safety signals in THIN, prior to the initial signal at the EMA. In many cases, this was because of the drug not being reliably captured in primary care data, and on a few occasions to the drugs not having yet been marketed in the UK. In contrast, some of the ADRs could be detected in VigiBase, even when the analysis was restricted to spontaneous reports from the UK. This shows that comprehensive surveillance for early safety signals requires broad population coverage as well as effective ascertainment of a wide spectrum of newly marketed drugs and adverse events. Concurrent research has found that even networks of longitudinal observational databases can be underpowered for rare adverse reactions, whereas common adverse reactions should be possible to detect for commonly used drugs [85]. There is an increasing number of observational databases available throughout the world for potential pharmacoepidemiology and signal detection work with each having widely varying characteristics, data structure and data quality concerns. It would be necessary to repeat the PROTECT analyses in these databases to determine to what extent these findings are generalisable.

Before approval of a drug, information on adverse events from clinical trials constitutes the primary basis for safety analysis and signal detection. PROTECT explored two statistical approaches to enhancing signal detection in clinical trials. One study explored the utility of extreme value modelling in early clinical studies as the basis for predicting drug toxicity in the subsequent phases of clinical development and evaluation [74]. A retrospective analysis showed that extreme value analysis of phase II data would have highlighted the risk for liver toxicity for a compound eventually withdrawn from the market on account of this risk. A second study evaluating different approaches to adjust for multiplicity found that Bayesian Hierarchical Models can improve signal detection performance through borrowing strength from related adverse events in the clinical trial dataset [75].

4 Conclusions

Over a period of 5 years, IMI PROTECT has addressed key research questions relevant to the science of safety signal detection. The resultant recommendations point to pragmatic steps that those working in the pharmacovigilance community can take to improve signal detection practices, whether in a national or international agency or a pharmaceutical company setting. PROTECT has also pointed to areas of potentially fruitful future research and some areas where further effort is likely to yield less.

References

1. Patwary KM. Report on statistical aspects of the pilot research project for international drug monitoring. Geneva: World Health Organisation; 1969.
2. Mandel SP, Levine A, Beleno GE. Signalling increases in reporting in international monitoring of adverse reactions to therapeutic drugs. Methods Inf Med. 1976;15:1–10.
3. Inman WH. Monitoring of adverse reactions to drugs in the United Kingdom. Proc R Soc Med. 1970;63:1302–4.
4. Napke E, Bishop J. The Canadian drug adverse reaction reporting program. Can Med Assoc J. 1966;95:1307.
5. Finney DJ. Statisical logic in the monitoring of reactions to therapeutic drugs. Methods Inf Med. 1971;10:237–45.
6. Hauben M, Noren GN. Editorial: a decade of data mining and still counting. Drug Saf. 2010;33:527–34.
7. CIOMS Working Group VIII. Practical aspects of signal detection in pharmacovigilance. 2010.
8. Pearson RK, Hauben M, Goldsmith DI, Gould AL, Madigan D, O'Hara DJ, Reisinger SJ, Hochberg AM. Influence of the MedDRA hierarchy on pharmacovigilance data mining results. Int J Med Inform. 2009;78:e97–103.
9. Hill R, Hopstadius J, Lerch M, Noren GN. An attempt to expedite signal detection by grouping related adverse reaction terms. Drug Saf. 2012;35:1194–5.
10. Trifiro G, Pariente A, Coloma PM, Kors JA, Polimeni G, Miremont-Salame G, Catania MA, Salvo F, David A, Moore N, Caputi AP, Sturkenboom M, Molokhia M, Hippisley-Cox J, Acedo CD, van der Lei J, Fourrier-Reglat A, EU-ADR group. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiol Drug Saf. 2009;18:1176–84.
11. Alvarez Y, Hidalgo A, Maignen F, Slattery J. Validation of statistical signal detection procedures in eudra vigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf. 2010;33:475–87.
12. Noren GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res. 2013;22:57–69.
13. Trombert-Paviot B, Rodrigues JM, Rogers JE, Baud R, Van Der Haring E, Rassinoux AM, Abrial V, Clavel L, Idir H. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. Int J Med Inform. 2000;58–59:71–85.
14. Tudorache T, Falconer S, Nyulas C, Storey MA, Ustun TB, Musen MA. Supporting the Collaborative Authoring of ICD-11 with WebProtege. In: AMIA Annual Symposium Proceedings; 2010. p. 802–806.
15. Bousquet C, Lagier G, Lillo-Le Louet A, Le Beller C, Venot A, Jaulent M. Appraisal of the MedDRA conceptual structure for describing and grouping adverse drug reactions. Drug Saf. 2005;28:19–34.
16. Souvignet J, Declerck G, Jaulent M, Bousquet C. Evaluation of automated term groupings for detecting upper gastrointestinal bleeding signals for drugs. Drug Saf. 2012;35:1195–6.

17. Souvignet J, Declerck G, Trombert B, Rodrigues JM, Jaulent MC, Bousquet C. Evaluation of automated term groupings for detecting anaphylactic shock signals for drugs. In: AMIA Annual Symposium Proceedings; 2012. p. 882–890.

18. Bousquet C, Sadou E, Souvignet J, Jaulent MC, Declerck G. Formalizing MedDRA to support semantic reasoning on adverse drug reaction terms. J Biomed Inform. 2014;49:282–91.

19. Declerck G, Bousquet C, Jaulent MC. Automatic generation of MedDRA terms groupings using an ontology. Stud Health Technol Inform. 2012;180:73–7.

20. European Union. Notice to applicants A guideline on summary of product characteristics (SmPC) revision 2. 2009. http://ec.europa.eu/health/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf. Accessed 15 Aug 2015.

21. Bergvall T, Dahlberg G, Cappelli B, Norén G. Fuzzy text matching to identify known adverse drug reactions. Pharmacoepidemol Drug Saf. 2011;20(S1):S143. doi:10.1002/pds.2206.

22. Hauben M, Reich L, Chung S. Postmarketing surveillance of potentially fatal reactions to oncology drugs: potential utility of two signal-detection algorithms. Eur J Clin Pharmacol. 2004;60:747–50.

23. European Medicines Agency. Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system. http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/11/WC500011437.pdf. Accessed 31 Jan 2016.

24. Boyce RD, Ryan PB, Noren GN, Schuemie MJ, Reich C, Duke J, Tatonetti NP, Trifiro G, Harpaz R, Overhage JM, Hartzema AG, Khayter M, Voss EA, Lambert CG, Huser V, Dumontier M. Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. Drug Saf. 2014;37:557–67.

25. IMI PROTECT ADR Database. http://www.imi-protect.eu/methodsRep.shtml. Accessed 17 Mar 2014.

26. EMA. Database structure. EMA. 2012. http://www.imi-protect.eu/documents/Databasestructure.pdf?_sm_au_=i2VfNqLvSJjQsqsH. Accessed 18 Jan 2015.

27. EMA EudraLex. Volume 2: pharmaceutical legislation notice to applicants and regulatory guidelines medicinal products for human use. European Commission. http://ec.europa.eu/health/documents/eudralex/vol-2/index_en.htm. Accessed 18 Jan 2015.

28. Slattery J, Cappelli B, Bergvall T, Opitz N, Kurz X. A structured database of adverse drug reaction based on information from the summary of product characteristics. Drug Saf. 2013;35:1192–3.

29. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54:315–21.

30. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53:177–90.

31. Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10:483–6.

32. Bate A, Evans SJW. Quantitative signal detection using spontaneous ADR reporting. Pharmacoepidemiol Drug Saf. 2009;18:427–36.

33. Van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection is spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11:3–10.

34. Gipson G. A shrinkage-based comparative assessment of observed-to-expected disproportionality measures. Pharmacoepidemiol Drug Saf. 2012;21:589–96.

35. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, Wisniewski A, Slattery J. Comparison of statistical signal detection methods within and across spontaneous reporting databases. Drug Saf. 2015;38:577–87.

36. Hopstadius J, Noren GN, Bate A, Edwards IR. Impact of stratification on adverse drug reaction surveillance. Drug Saf. 2008;31:1035–48.

37. Caster O, Juhlin K, Watson S, Norén GN. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank: retrospective evaluation against emerging safety signals. Drug Saf. 2014;37:617–28.

38. Van Holle L, Bauchau V. Optimization of a quantitative signal detection algorithm for spontaneous reports of adverse events post immunization. Pharmacoepidemiol Drug Saf. 2013;22:477–87.

39. Hopstadius J, Norén GN. Robust discovery of local patterns: subsets and stratification in adverse drug reaction surveillance. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium; 2012. p. 265–274.

40. Woo EJ, Ball R, Burwen DR, Braun MM. Effects of stratification on data mining in the US Vaccine Adverse Event Reporting System (VAERS). Drug Saf. 2008;31:667–74.

41. Zeinoun Z, Seifert H, Verstraeten T. Quantitative signal detection for vaccines: effects of stratification, background and masking on GlaxoSmithKline's spontaneous reports database. Hum Vaccines. 2009;5:599–607.

42. Almenoff JS, LaCroix KK, Yuen NA, Fram D, DuMouchel W. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf. 2006;29:875–87.

43. Grundmark B, Holmberg L, Garmo H, Zethelius B. Reducing the noise in signal detection of adverse drug reactions by standardizing the background: a pilot study on analyses of proportional reporting ratios-by-therapeutic area. Eur J Clin Pharmacol. 2014;70:627–35.

44. De Bie S, Verhamme KMC, Straus SMJM, Stricker BHC, Sturkenboom MCJM. Vaccine-based subgroup analysis in Vigibase: effect on sensitivity in paediatric signal detection. Drug Saf. 2012;35:335–46.

45. Seabroke S, Candore G, Juhlin K, Quarcoo N, Wisniewski A, Arani R, Painter J, Tregunno P, Norén GN, Slattery J. Performance of stratified and subgrouped disproportionality analyses in spontaneous databases. Drug Saf. 2016. doi:10.1007/s40264-015-0388-3.

46. Pariente A, Didailler M, Avillach P, Miremont-Salame G, Fourrier-Reglat A, Haramburu F, Moore N. A potential competition bias in the detection of safety signals from spontaneous reporting databases. Pharmacoepidemiol Drug Saf. 2010;19:1166–71.

47. Pariente A, Avillach P, Salvo F, Thiessard F, Miremont-Salam G, Fourrier-Reglat A, Haramburu F, Bgaud B, Moore N. Effect of competition bias in safety signal generation: analysis of a research database of spontaneous reports in France. Drug Saf. 2012;35:855–64.

48. Juhlin K, Ye X, Star K, Noren GN. Outlier removal to uncover patterns in adverse drug reaction surveillance: a simple unmasking strategy. Pharmacoepidemiol Drug Saf. 2013;22:1119–29.

49. Maignen F, Hauben M, Hung E, Holle LV, Dogne J-. A conceptual approach to the masking effect of measures of disproportionality. Pharmacoepidemiol Drug Saf. 2014;23:208–17.

50. Maignen F, Hauben M, Hung E, Van Holle L, Dogne J. Assessing the extent and impact of the masking effect of disproportionality analyses on two spontaneous reporting systems databases. Pharmacoepidemiol Drug Saf. 2014;23:195–207.

51. Wang H, Hochberg AM, Pearson RK, Hauben M. An experimental investigation of masking in the US FDA adverse event reporting system database. Drug Saf. 2010;33:1117–33.

52. Noren GN, Sundberg R, Bate A, Edwards IR. A statistical methodology for drug-drug interaction surveillance. Stat Med. 2008;27:3057–70.

53. Thakrar BT, Grundschober SB, Doessegger L. Detecting signals of drug-drug interactions in a spontaneous reports database. Br J Clin Pharmacol. 2007;64:489–95.

54. Strandell J, Caster O, Hopstadius J, Edwards IR, Noren GN. The development and evaluation of triage algorithms for early discovery of adverse drug interactions. Drug Saf. 2013;36:371–88.

55. Soeria-Atmadja D, Juhlin K, Thakrar B, Norén G. Evaluation of statistical measures for adverse drug interaction surveillance. Pharmacoepidemiol Drug Saf. 2014;23(S1):294–5.

56. Hauben M, Reich L, DeMicco J, Kim K. 'Extreme duplication' in the US FDA adverse events reporting system database. Drug Saf. 2007;30:551–4.

57. Tregunno PM, Fink DB, Fernandez-Fernandez C, Lazaro-Bengoa E, Noren GN. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf. 2014;37:249–58.

58. Noren GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14:305–28.

59. Maciá-Martínez M, de Abajo FJ, Roberts G, Slattery J, Thakrar B, Wisniewski AFZ. An empirical approach to explore the relationship between measures of disproportionate reporting and relative risks from analytical studies. Drug Saf. 2015. doi:10.1007/s40264-015-0351-3.

60. Evans SJW. What Is the plural of a 'yellow' anecdote? Drug Saf. 2015. doi:10.1007/s40264-015-0368-7.

61. Norén GN, Hopstadius J, Bate A, Star K, Edwards IR. Temporal pattern discovery in longitudinal electronic patient records. Data Min Knowl Discov. 2010;20:361–87. doi:10.1007/s10618-009-0152-3.

62. Choi N, Chang Y, Choi YK, Hahn S, Park B. Signal detection of rosuvastatin compared to other statins: data-mining study using national health insurance claims database. Pharmacoepidemiol Drug Saf. 2010;19:238–46.

63. Dore DD, Seeger JD, Chan KA. Use of a claims-based active drug safety surveillance system to assess the risk of acute pancreatitis with exenatide or sitagliptin compared to metformin or glyburide. Curr Med Res Opin. 2009;25:1019–27.

64. Curtis JR, Cheng H, Delzell E, Fram D, Kilgore M, Saag K, Yun H, Dumouchel W. Adaptation of bayesian data mining algorithms to longitudinal claims data: Coxib safety as an example. Med Care. 2008;46:969–75.

65. Jin HW, Chen J, He H, Williams GJ, Kelman C, O'Keefe CM. Mining unexpected temporal associations: applications in detecting adverse drug reactions. IEEE Trans Inf Technol Biomed. 2008;12:488–500.

66. Noren GN, Hopstadius J, Bate A, Edwards IR. Safety surveillance of longitudinal databases: methodological considerations. Pharmacoepidemiol Drug Saf. 2011;20:714–7.

67. Schuemie MJ. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. Pharmacoepidemiol Drug Saf. 2011;20:292–9.

68. Walker AM. Signal detection for vaccine side effects that have not been specified in advance. Pharmacoepidemiol Drug Saf. 2010;19:311–7.

69. Moore TJ, Furberg CD. Electronic health data for postmarket surveillance: a vision not realized. Drug Saf. 2015;38:601–10.

70. Bate A, Brown EG, Goldman SA, Hauben M. Terminological challenges in safety surveillance. Drug Saf. 2012;35:79–84.

71. Edwards I, Karimi G, Bergvall T, Caster O, Asiimwe A, Star K, Bhayat F, Soriano-Gabarro M, Thakrar B, Thompson M, Bate A, Norén G. Risk identification in healthcare records: comparison to epidemiological studies. Pharmacoepidemiol Drug Saf. 2013;22(S1):311–2.

72. Cederholm S, Hill G, Asiimwe A, Bate A, Bhayat F, Persson Brobert G, Bergvall T, Ansell D, Star K, Norén GN. Structured assessment for prospective identification of safety signals in electronic medical records: evaluation in the health improvement network. Drug Saf. 2015;38:87–100.

73. Trifiro G, Patadia V, Schuemie MJ, Coloma PM, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Scotti L, Pedersen L, Avillach P, Sturkenboom MC, van der Lei J. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. Stud Health Technol Inform. 2011;166:25–30.

74. Southworth H. Predicting potential liver toxicity from phase 2 data: a case study with ximelagatran. Stat Med. 2014;33:2914–23.

75. Ahlers C, Brueckner A, Mallick A, Opitz N, Pinkston V, Tran B, Scott J, Southworth H, Van Holle L, Wallis N. Statistical signal detection in Clinical Trial data. http://www.imi-protect.eu/documents/PROTECTsymposium_signaldetectioninClinicaltrials_talk.pdf. Accessed 12 Dec 2015.

76. Southworth H, Heffernan JE. Multivariate extreme value modelling of laboratory safety data from clinical studies. Pharm Stat. 2012;11:367–72.

77. European Medicines Agency (EMA) Guideline on good pharmacovigilance practices (GVP) Module IX—signal management. 2012.

78. Wisniewski AFZ, Juhlin K, Laursen M, Macia MM, Manlik K, Pinkston VK, Seabroke S, Slattery J. Characterisation of databases (DBS) used for signal detection (SD): results of a survey of imi protect work package (WP) 3 participants. Pharmacoepidemiol Drug Saf. 2012;21:233–4.

79. Ryan PB, Madigan D, Stang PE, Marc-Overhage J, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012;31:4401–15.

80. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, Van Der Lei J, Sturkenboom M. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf. 1011;20:1–11.

81. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. Ann Intern Med. 2010;153:600–6.

82. Schuemie MJ, Gini R, Coloma PM, Straatman H, Herings RMC, Pedersen L, Innocenti F, Mazzaglia G, Picelli G, Van Der Lei J, Sturkenboom MCJM. Replication of the OMOP experiment in europe: evaluating methods for risk identification in electronic health record databases. Drug Saf. 2013;36(S159):S169.

83. Schuemie MJ, Coloma PM, Straatman H, Herings RMC, Trifiro G, Matthews JN, Prieto-Merino D, Molokhia M, Pedersen L, Gini R, Innocenti F, Mazzaglia G, Picelli G, Scotti L, Van Der Lei J, Sturkenboom MCJM. Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med Care. 2012;50:890–7.

84. Harpaz R, Vilar S, DuMouchel W, Salmasian H, Haerian K, Shah NH, Chase HS, Friedman C. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc. 2013;20:413–9.

85. Coloma PM, Trifirò G, Schuemie MJ, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Picelli G, Corrao G, Pedersen L, van der Lei J, Sturkenboom M. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? Pharmacoepidemiol Drug Saf. 2012;21:611–21.