ORIGINAL RESEARCH ARTICLE

Drug Saf 2010; 33 (12): 1117-1133
0114-5916/10/0012-1117/$49.95/0

An Experimental Investigation of Masking in the US FDA Adverse Event Reporting System Database

Hsin-wei Wang,[1] *Alan M. Hochberg*,[2] *Ronald K. Pearson*[2] and *Manfred Hauben*[3,4,5,6]

1 Computational Sciences Centre of Emphasis, Pfizer Global Research and Development, Cambridge, Massachusetts, USA
2 ProSanos Corporation, Harrisburg, Pennyslvania, USA
3 Risk Management Strategy, Pfizer Inc., New York, New York, USA
4 New York University School of Medicine, New York, New York, USA
5 New York Medical College, Valhalla, New York, USA
6 Brunel University, West London, UK

Abstract

Background: A phenomenon of 'masking' or 'cloaking' in pharmacovigilance data mining has been described, which can potentially cause signals of disproportionate reporting (SDRs) to be missed, particularly in pharmaceutical company databases. Masking has been predicted theoretically, observed anecdotally or studied to a limited extent in both pharmaceutical company and health authority databases, but no previous publication systematically assesses its occurrence in a large health authority database.

Objective: To explore the nature, extent and possible consequences of masking in the US FDA Adverse Event Reporting System (AERS) database by applying various experimental unmasking protocols to a set of drugs and events representing realistic pharmacovigilance analysis conditions.

Methods: This study employed AERS data from 2001 through 2005. For a set of 63 Medical Dictionary for Regulatory Activities (MedDRA®) Preferred Terms (PTs), disproportionality analysis was carried out with respect to all drugs included in the AERS database, using a previously described urn-model-based algorithm. We specifically sought masking in which drug removal induced an increase in the statistical representation of a drug-event combination (DEC) that resulted in the emergence of a new SDR. We performed a series of unmasking experiments selecting drugs for removal using rational statistical decision rules based on the requirement of a reporting ratio (RR) >1, top-ranked statistical unexpectedness (SU) and relatedness as reflected in the WHO Anatomical Therapeutic Chemical level 4 (ATC4) grouping. In order to assess the possible extent of residual masking we performed two supplemental purely empirical analyses on a limited subset of data. This entailed testing every drug and drug group to determine which was

most influential in uncovering masked SDRs. We assessed the strength of external evidence for a causal association for a small number of masked SDRs involving a subset of 29 drugs for which level of evidence adjudication was available from a previous study.

Results: The original disproportionality analysis identified 8719 SDRs for the 63 PTs. The SU-based unmasking protocols generated variable numbers of masked SDRs ranging from 38 to 156, representing a 0.43–1.8% increase over the number of baseline SDRs. A significant number of baseline SDRs were also lost in the course of our experiments. The trend in the number of gained SDRs per report removed was inversely related to the number of lost SDRs per protocol. Both the number and nature of the reports removed influenced the number of gained SDRs observed. The purely empirical protocols unmasked up to ten times as many SDRs. None of the masked SDRs had strong external evidence supporting a causal association. Most involved associations for which there was no external supporting evidence or were in the original product label. For two masked SDRs, there was external evidence of a possible causal association.

Conclusions: We documented masking in the FDA AERS database. Attempts at unmasking SDRs using practically implementable protocols produced only small changes in the output of SDRs in our analysis. This is undoubtedly related to the large size and diversity of the database, but the complex interdependencies between drugs and events in authentic spontaneous reporting system (SRS) databases, and the impact of measures of statistical variability that are typically used in real-world disproportionality analysis, may be additional factors that constrain the discovery of masked SDRs and which may also operate in pharmaceutical company databases. Empirical determination of the most influential drugs may uncover significantly more SDRs than protocols based on predetermined statistical selection rules but are impractical except possibly for evaluating specific events. Routine global exercises to elicit masking, especially in large health authority databases are not justified based on results available to date. Exercises to elicit unmasking should be driven by prior knowledge or obvious data imbalances.

Background

'Masking' or 'cloaking' is a term used in pharmacovigilance data mining to denote the suppression of a statistical reporting association between a drug and an adverse event due to large numbers of reports for that adverse event in connection with another drug or drugs. When one drug has a disproportionately large number of reports for a particular adverse event, it makes that event appear more common. This tends to make occurrences of that event for another drug appear statistically less unusual. Masking is mentioned in a number of descriptive articles and guidances on data mining techniques.[1-4] It is described clearly by Gould,[1] who gives an example from a company safety database, using two anonymous related drugs. Masking is normally discussed within the context of manufacturer databases, where report tallies may be influenced by closely related pairs

or groups of drugs, rather than for more diverse databases, such as the US FDA Adverse Event Reporting System (AERS) database. However, previous publications do not rule out the possibility of masking in these larger and more diverse databases and in fact state that it may occur in any database.[2] Currently available guidances and white papers on data mining in pharmacovigilance do not make blanket recommendations on how to account for masking in pharmacovigilance analysis.[2,3] Limited and non-systematic analysis have looked for masking with drugs in the FDA AERS database[5] and with vaccines in a company database,[6] but presently it is more of a theoretical concern.

The 'classical' formulation of masking requires that a statistical association be suppressed to some degree.[1] In the most common application scenario in which drug safety data mining is used for binary classification based on a threshold of disproportionality, masking might be considered significant if the magnitude of the unmasked association increases beyond the threshold defining a signal of disproportionate reporting (SDR).[7] This would depend on multiple factors, including not just the extent of reporting of the event of interest with the potentially masking drug, but also on how near to the SDR-defining threshold the potentially masked association is to begin with. For example, in a data mining exercise with an SDR-defining threshold of 2.0, an initial association with a reporting ratio (RR) of 1.9 might be more likely to become 'unmasked' (an increase in RR of 0.1 or 5.2%) compared with an association with an initial RR of 0.5 (an increase of 1.5 or 300%). A detailed numerical explanation of masking is included in the Appendix.

The possibility of masking raises the question of whether statistical association measures should be recomputed after selective removal of targeted subsets of reports and, if so, which subset of reports to remove. We might re-compute statistical measures of association for a drug of interest, 'Drug A', after first removing from the database a set of reports for another drug, 'Drug B', that is suspected of masking one or more of Drug A's SDRs. This approach assumes that there is an *a priori* suspicion that masking of Drug A's statistical reporting associations may be taking place, and an *a priori* reason to suspect that the related drug, Drug B, is responsible for it. In the example of Gould,[1] Drug A and Drug B are two compounds with a similar mechanism of action produced by the same manufacturer. Under these conditions, one can readily conduct a secondary analysis, implementing the strategy outlined by Gould[1] to deal with potential masking in a pharmaceutical company database. An important caveat is that published arithmetic explanations of masking do not accurately reflect real-world pharmacovigilance scenarios in two respects. They are based on calculated disproportionality measures, ignoring the influence of measures of statistical significance, which are typically used in parallel with disproportionality measures in pharmacovigilance, and they involve an idealized database in which each report lists a single suspect drug.[1,2]

Pharmacovigilance databases are often employed to screen large numbers of drugs for potential signals for large numbers of events. It is unclear under these circumstances how one would expeditiously decide what constitutes the 'related' drug, or set of drugs, for each 'Drug A' to be analysed, and what impact, if any, their elimination might have on overall signal detection performance. Should we seek out 'Drug B' candidates, and conduct secondary analyses for masking, whenever we conduct data mining? Should such a procedure be routinized based on rational statistical decision rules for drug removal or should such secondary analyses be limited to those cases where there is a high prior suspicion of masking? If primary and secondary analyses yield discordant results, what criteria are used to adjudicate between them or arrive at a unified interpretation? These questions take on additional importance due to the current lack of a standardized procedure for eliciting masking and the warning against selectively retrofitting an analysis to pre-existing biases or expectations.[8]

Since masking has traditionally been discussed as occurring with single drugs in pharmaceutical company databases, it seems an intuitively plausible extension that if masking exists in larger public databases it may possibly be elicited if the

aforementioned strategy of single drug removal is replaced with removal of larger numbers of drugs and/or drug groups. For example, a possible extension of the unmasking strategy described by Gould,[1] whose formulation of masking invokes pharmacological relatedness between drugs, would be to remove multiple drugs from within a chemically or pharmacologically related group that are strongly associated with the event of interest, and check for unmasked SDRs with the remaining chemically related drugs.

It would be helpful to know whether masking is a generally prevalent and consequential phenomenon in pharmacovigilance data mining, or whether it is relatively rare and unimportant and how it may vary as a function of the unmasking strategy. If undetected masking of SDRs occurs frequently, and significantly affects our signal detection capability, then perhaps a proactive approach to the problem is justified, and we should change the way that we routinely data mine to account for masking. If masking is a rare phenomenon, then we may only need to account for it on an *ad hoc* basis when we are studying events known to be strongly associated with a particular small set of drugs.

Herein we report on the nature and frequency of masking observed in the AERS database with the aforementioned strategy of removing of multiple drugs/drug groups.

Methods

Data Sources

Data for the study was taken from the public release, known as the "Freedom of Information Act (FOI)" version, of the FDA AERS database, covering the period from the first quarter of 2001 through the end of 2005. In the AERS database, adverse events are described at the Medical Dictionary for Regulatory Activities (MedDRA®)[1] Preferred Term (PT) level. Obsolete MedDRA® PTs in AERS were updated to Version 10.1.

Masking Events

To identify 'Drug B' candidates, we first created a list of *events*, corresponding to the 'Event mentioned' in the Appendix example. We refer to these events as *masking events*. Our list of 63 masking events consisted of two parts. The first was a set of 47 adverse events selected at random from a list of all events that gave SDRs for one or more of the drugs in a previous study of drug safety data mining.[9] The second was a convenience sample of 16 adverse events chosen by a safety physician (MH) as representing some of the more common events of concern in pharmacovigilance. They were not chosen with regard to subjective concern, or lack of concern, related to masking. The full list of masking events is shown in table I.

Strategy to Identify Masking

Fundamental to uncovering masked SDRs for a given event is to identify and remove from the database all reports involving drugs whose statistical reporting association with that event might be responsible for masking. A strategy to identify masking must strike a balance between thoroughness and practicality both for research purposes and real-world pharmacovigilance. For example, one approach would be for a given event to remove every drug in the database one at a time, empirically testing which drug, if removed, elicits the most masked SDRs, and then repeat this procedure with each iteration of the protocol. While this might elicit the most masked SDRs, it is computationally intractable for a full analysis (we estimated that using this approach would have required 6.5 years to complete for our entire sample of drugs and events).

In contrast to selecting drugs for removal based purely on empirical testing, we used rational statistical criteria to choose which drugs to remove. When selecting a criterion for drug removal, two readily assessable factors that can influence the masking potential of a drug for a specific event are the RR and the number of reports of the drug-event combination (DEC) [Nab]. DECs

1 MedDRA® is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

with a very high RR and Nab have the greatest masking potential. Drugs with lower RR and lower Nab have less potential to mask SDRs. In between there will be different effects of various combinations of higher/lower RR and Nab. Therefore we only considered removing associations with an RR >1, which is a pre-requisite for masking as described by Gould,[1] but we then used the Statistical Unexpectedness (SU) value from disproportionality analysis as the basis for further measuring the strength of association because it incorporates both disproportionality information and also is influenced by case count. We examined the RR alone to identify masking candidates, but found it unsuitable for our experimental purposes. As the equations by Gould[1] make clear, the ability of a 'Drug B' to influence results for 'Drug A' depends directly on the number of reports (designated 'n_{21}' in Gould's notation), which mention Drug B in association with the event in question. The top RR for each event was often a drug that appears rarely in the AERS database, sometimes in only one or two individual case safety reports (ICSRs).

Specifically, we used four computationally tractable and variably intensive strategies for eliciting masked SDRs based on the requirement of the drug/drug grouping having an RR >1 and the strongest statistical association based on top-ranked SU. Taking each of the 63 events of our study in turn, our strategy was to experimentally remove the following from the AERS database:

1. The WHO Anatomical Therapeutic Chemical (ATC) level 4 (ATC4) drug group containing the single drug that was most strongly associated with the event (Top-Drug-ATC4). Because the drugs within a given ATC4 drug group have therapeutic, pharmacological and chemical similarity, we considered this protocol most analogous to the classical formulation of masking that invokes mechanistic relatedness between drugs,[1] but scaled-up to a health authority database.
2. The ATC4 drug class that was most strongly associated with the event (Top-ATC4-SU).
3. The ten drugs with the strongest statistical association (10-Drug-SU). In making the latter determination for each of the 63 PTs, we repeated disproportionality analysis and ranking after each

Table I. MedDRA® Preferred Terms (PTs) studied as masking events

Random PTs	Convenience PTs[a]
Aggression	Anaphylactic reaction
Agitation	Blood glucose increased
Anaemia	Bradycardia
Back disorder	Convulsion
Bacterial infection	Cough
Blood potassium increased	Dizziness
Bone marrow failure	Dysgeusia
Bronchopulmonary aspergillosis	Gastrointestinal haemorrhage
Cardiac failure congestive	Headache
Cellulitis	Hyperglycaemia
Clostridial infection	Hypertension
Confusional state	Hyponatraemia
Coordination abnormal	Neutropenia
Diabetes mellitus non-insulin-dependent	Pancreatitis
Diplopia	Thrombocytopenia
Dysarthria	Urticaria
Dyskinesia	
Dysstasia	
Endophthalmitis	
Erectile dysfunction	
Euphoric mood	
Extrapyramidal disorder	
Eye irritation	
Fluid overload	
Gingival hyperplasia	
Hallucination	
Hepatotoxicity	
Hypophosphataemia	
Intraocular pressure increased	
Left ventricular hypertrophy	
Lipase increased	
Liver function test abnormal	
Myalgia	
Myoclonus	
Nausea	
Oedema peripheral	
Oral pain	
Palmar-plantar erythrodysaesthesia syndrome	
Panic reaction	
Peritonitis	

Continued next page

Table I. Contd

Random PTs	Convenience PTs[a]
Pneumonitis	
Rash erythematous	
Rash generalized	
Rash maculo-papular	
Renal failure	
Ventricular tachycardia	
White blood cell count decreased	

a Convenience samples are 16 adverse events chosen by a safety physician (MH) as representing some of the more common events of concern in pharmacovigilance.

round of drug elimination. While rankings generally remained stable, some drugs significantly dropped in rank as others were eliminated. We therefore subtracted the *single* most strongly associated drug from each of *ten* sequential disproportionality analyses, rather than the top ten most strongly associated drugs from a *single* disproportionality analysis. We summed the number of unique SDRs gained at each step.

4. The top ten drugs ranked by the strength of the statistical association in the original dataset prior to removing any drugs (10-Drug-1-Step). In other words, we removed en bloc reports of the first, second, third … tenth most strongly associated drugs for each event.

Finally, we estimated the extent to which our statistical criteria may have missed residual masked SDRs by performing a purely empirical determination (irrespective of RR and SU) of which drugs generate the most masked SDRs when removed from the dataset. We selected a single event, myoclonus, and empirically determined which drug and which ATC4 group were each most influential in generating masked SDRs for this event when removed from the dataset (10-Drug-Empirical and Top-ATC4-Empirical, respectively). For the former, we removed that drug and then repeated the analysis until a total of ten drugs (ties for the tenth drug resolved with random selection) were removed that represented the drugs most influential in eliciting masked SDRs. We compared the number of SDRs generated by these empirical protocols to the corresponding SU-based protocols.

To assist the reader in following the results for the multiple protocols, table II provides a convenient reference key of protocol names and implementation details.

For all of the above protocols, we adhered to a list of invalid drugs that were not considered to be candidates for removal due to factors such as lack of specific/meaningful drug nomenclature or lack of biological plausibility.

Disproportionality Analysis

Disproportionality analysis was carried out using the urn-model method, which is described, along with the specific implementation details used for this analysis, by Hochberg et al.[9] Each data mining run for a particular drug produced a list of drug-MedDRA® PT pairs, which were considered SDRs. For each SDR, we tabulated the number of ICSRs for the DEC, the RR and the SU. We calculated the median and corre-

Table II. Reference key to unmasking protocols

Protocol	Methodology
Top-Drug-ATC4	1. Identify ATC4 group with the single drug having the strongest statistical association[a] with the event being studied 2. Remove all drugs within that ATC4 that have RR >1
Top-ATC4-SU	Remove ATC4 drug group with the strongest statistical association[a] with the event being studied
10-Drug-SU	Remove the drug with the strongest statistical association[a] and repeat with the remaining data for a total of ten iterations summing unique SDRs gained in each step
10-Drug-1-Step	Remove en bloc the ten drugs with the strongest statistical association[a] (i.e. remove the first, second, third … tenth most strongly associated drug)
10-Drug-Empirical (myoclonus[b])	Test every drug in the database and determine which would generate the most masked SDRs if removed, and repeat ten times, summing unique SDRs gained in each step
Top-ATC4-Empirical (myoclonus[b])	Test every ATC4 to determine which would generate the most masked SDRs, and remove it

a RR >1 and top-ranked SU.

b The event myoclonus was chosen as a representative example adverse event for illustrative purposes.

ATC4 = Anatomical Therapeutic Chemical level 4; **RR** = reporting ratio; **SDRs** = signals of disproportionate reporting; **SU** = statistical unexpectedness.

sponding lower and upper quartiles of the RR and SU of the drugs/drug groups removed per protocol, and the number of drugs and number of reports removed per protocol. We used medians and lower and upper quartiles as descriptive statistics for these parameters because they are non-negative values with a strongly skewed distribution.

For each protocol we calculated the number of unique SDRs gained for each event. The qualifier 'unique' is applicable to the protocols involving iterative (i.e. ten drug/step) analysis. In these protocols it is possible for a masked SDR to be uncovered and then remasked/unmasked one or more times during sequential iterations. A masked SDR was counted once it first appeared. On the theory that SDRs that are lost during drug removal may be a 'signal' that some masked SDRs are being further masked, we also calculated the number of SDRs lost per protocol. To gain insight into the mechanism of lost SDRs we calculated the number of SDRs lost due to a drop below SDR-defining thresholds of the RR, the SU or both.

Assessment of Relationship Between Drugs in Signals of Disproportionate Reporting (SDRs) and Removed Drugs

We used the WHO ATC Classification System hierarchy to assess whether the drug in each SDR was related to the drug that had been removed from the data to stimulate masking. The ATC4 drug group was considered to be a reasonable surrogate for 'relatedness' as described in the classic formulation of masking since it defines relevant chemical structure related to a given therapeutic/pharmacological activity. This initial choice was supported by a limited qualitative review of ATC drug hierarchy assignments by one of the investigators (MH).

Importance of Masked SDRs

Masked SDRs may tend to be close to the detection threshold, raising the question of their practical importance in pharmacovigilance. For a subset of 29 drugs that were analysed in a previous study, we classified the masked SDRs according to a level of evidence scheme that has been fully described previously in *Drug Safety*.[10] Briefly, unmasked SDRs were classified as in the original label – definite, probable, possible or unsupported by evidence based on multiple information sources.

Results

Baseline Analysis

Coverage by the 63 MedDRA® Preferred Terms

We have calculated the proportion of reports covered by the use of 63 MedDRA® PTs. The dataset taken from AERS between 2001 and 2005 contained 898 566 ICSRs, once initial reports and follow-up reports were 'rolled up' into a single ICSR. At least one of our 63 events is mentioned in 297 299 of these. This total refers to numbers of ICSRs and not to numbers of DECs; if two of our 63 events occur in a single ICSR, it is only counted once. On this basis, our 63 events are involved in 33.09% of ICSRs.

Drugs that were eliminated had, on average, more reports for the masking event than other drugs. For example, drugs removed in the 10-Drug-SU protocol are associated with a median of 104.0 (quartile [Q] 3: 312.8; Q1: 34) reports versus 4.0 (17.0; 1.0) for other drugs. Results from the same protocol also show that eliminated drugs had a higher RR of 6.049 (Q3: 11.920; Q1: 3.852) versus 1.548 (2.948; 0.892) for other drugs. For other protocols please refer to figure 1.

Baseline SDRs

Using FOI AERS data from 2001 to 2005, we looked at all possible drug associations for the 63 MedDRA® PTs, using disproportionality analysis. We obtained a total of 8719 SDRs. There was a median of 69.0 ICSRs in each SDR (Q3: 189.0; Q1: 27.0). [These statistics represent 'cell *a*' in Gould's notation: counts of reports where both event and drug are present.] The median RR was 2.660 (Q3: 4.785; Q1: 1.859). The median SU was 9.732 (19.940; 6.098).

Characteristics of Drugs/Drug Groups Removed Per Protocol

The 10-Drug-Empirical was associated with the highest median number of reports removed

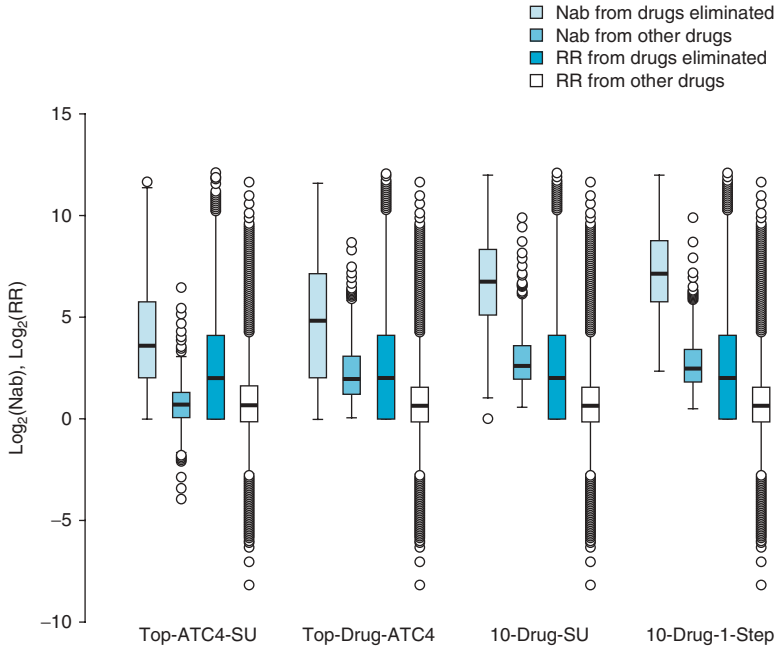

Fig. 1. Boxplot showing number of reports of the drug-event combination (Nab) and relative risk (RR) associated with drugs removed vs other drugs using the four protocols. **ATC4** = Anatomical Therapeutic Chemical level 4; **SU** = statistical unexpectedness.

(82 334) followed by the 10-Drug-1-Step (79 070), 10-Drug-SU (77 000), Top-ATC4-SU (69 740) and Top-Drug-ATC4 (25 530). Drugs removed in the 10-Drug-SU and 10-Drug-1-Step protocols had substantially higher SUs and involved a greater number of reports removed than the corresponding protocol involving removal of drugs at the ATC4 level. The vast majority of reports removed did not involve the PT being analysed (>98%) for all protocols.

The median and first and third quartiles for the RR, SU, number of drugs and number of reports for the drugs/drug groups removed in each protocol are displayed in table III.

Per Protocol Number of Gained SDRs

The 10-Drug-1-Step unmasking protocol resulted in a gain of 38 SDRs, representing a 0.43% gain over the 8719 baseline SDRs. The Top-Drug-ATC4 unmasking protocol resulted in a gain of 64 SDRs, representing a 0.73% increase over baseline. The 10-Drug-SU protocol generated 100 masked SDRs, representing a 1.1% increase over baseline. The Top-ATC4-SU unmasking protocol resulted in a gain of 156 SDRs, representing a 1.8% increase over baseline. Each protocol also resulted in significant numbers of lost SDRs (table IV).

The 10-Drug-empirical protocol generated ten masked SDRs for myoclonus compared with one generated by the 10-Drug-SU protocol. The Top-ATC4-empirical protocol generated six masked SDRs for myoclonus compared with five generated by the Top-ATC4-SU protocol.

Two mechanisms may explain lost SDRs. One is that reports of the drug removed per protocol lists additional drugs associated with an SDR with the same event, which are therefore simultaneously removed from the database with the per protocol removed drug and therefore no longer 'at risk' of being discovered as a masked SDR ('collateral removal'). Another mechanism is the change in the database size and/or background that occurs due to both per protocol and collateral drug removal for remaining drugs. The

majority of lost SDRs were due to the SU dropping below threshold (table IV).

Figures 2 and 3 display the number of gained and lost SDRs per protocol relative to the 10-Drug-1-Step, which had the fewest gained and the most lost SDRs. The gained and lost figures are presented as the raw totals and per median number of drugs (figure 2) and per median number of reports (figure 3) removed. There is a clear inverse relationship between the trend in the per protocol number of SDRs gained per report removed and the number lost.

Since the Top-Drug-ATC4 unmasking protocol is analogous to the classical formulation of masking invoking relatedness but scaled up to a large health authority database, we calculated there were 276 of the 8719 baseline SDRs in Top-Drug-ATC4 groups. None of the 64 SDRs generated by the Top-Drug-ATC4 protocol involved the same ATC4 group – expressed a little differently, none of the gained SDRs involved the remaining pharmacologically related drugs in the respective Top-Drug-ATC4 group that had an initial RR <1 and were therefore not removed per protocol. Therefore all gained SDRs involved drugs that were not chemically related to the removed drugs.

Statistical Unexpectedness-Based versus Empirical Drug Removal

The purely empirical protocols generated more SDRs and resulted in fewer lost SDRs than the SU-based protocols. However, this difference was only marked for the iterative ten-step protocols and not the protocols based on removal of ATC4 groups. Table V recapitulates the findings from the SU-based and empirical protocols.

Unmasked SDRs and Their Importance

Of the 38, 64, 100 and 156 SDRs unmasked by the four SU-based protocols based on statistical

Table III. Characteristics of drugs removed per protocol

Protocol	RR (Q3; Q1)	SU (Q3; Q1)	No. of drugs (Q3; Q1)		N (Q3; Q1)
Top-ATC4-SU	1.823 (2.252; 1.614)	14.530 (36.06; 7.059)	All	14.0 (22.5; 7.0)	69 740 (79 100; 37 580)
			Listing events	4.0 (7.5; 2.5)	185.0 (663.5; 53.5)
Top–Drug-ATC4	6.866 (43.260; 1.306)	3.936 (8.388; 2.296)	All	6.0 (8.5; 3.5)	25 530 (59 720; 13 180)
			Listing events	5.0 (8.0; 3.0)	400.0 (1450.0; 211.5)
10-Drug-SU	5.478 (10.68; 3.337)	29.510 (69.85; 12.05)	All	10.0 (10; 10)	77 000 (121 700; 51 200)
			Listing events	10.0 (10; 10)	1428.0 (4576; 471)
10-Drug-1-Step	5.546 (10.510; 3.444)	56.180 (127.60; 26.40)	All	10.0 (10; 10)	79 070 (151 900; 51 930)
			Listing events	10.0 (10; 10)	1698.0 (5308; 598.5)
10-Drug-Empirical (myoclonus[a])	2.882 (3.815; 2.646)	6.015 (7.262; 5.014)	All	10.0	82 334
			Myoclonus	10.0	311
Top-ATC4-Empirical (myoclonus[a])	1.169 (2.044; 0.909)	0.5 (0.9941; 0.1305)	All	7	61 994
			Myoclonus	5	68

a The event myoclonus was chosen as a representative example adverse event for illustrative purposes.

ATC4 = Anatomical Therapeutic Chemical level 4; **N** = number of reports for the drugs/drug groups removed in each protocol; **Q** = quartile; **RR** = reporting ratio; **SU** = statistical unexpectedness.

Table IV. Number of signals of disproportionate reporting (SDRs) gained and lost by masking protocol (baseline number of SDRs = 8719)

Protocol	SDRs gained	SDRs lost					
		lost by removal			lost by secondary masking[a]		
		total no.	method	no.	total no.	parameter	value
Top-ATC4-SU	156	101	On removal list	80	1787	RR	0
			Collateral removal[b]	21		RR and SU	15
						SU	1772
Top-Drug-ATC4	64	295	On removal list	218	2462	RR	0
			Collateral removal[b]	77		RR and SU	151
						SU	2311
10-Drug-SU	100	876	On removal list	630	5919	RR	0
			Collateral removal[b]	246		RR and SU	1266
						SU	4653
10-Drug-1-Step	38	946	On removal list	630	5454	RR	0
			Collateral removal[b]	316		RR and SU	1293
						SU	4161
10-Drug-Empirical (myoclonus[c])	10	11	On removal list	10	15	RR	0
			Collateral removal[b]	1		RR and SU	0
						SU	15
Top-ATC4-Empirical (myoclonus[c])	6	5[d]	On removal list	5	4	RR	0
			Collateral removal[b]	0		RR and SU	0
						SU	4

a Secondary masking refers to masking not induced by collateral removal for associations initially/previously associated with an SDR.

b Collateral removal means that all reports for that drug-event combination were simultaneously removed with per protocol drug.

c The event myoclonus was chosen as a representative example adverse event for illustrative purposes.

d ATC4 group A07EA (Corticosteroids Acting Locally) has seven members but five are associated with myoclonus.

ATC4 = Anatomical Therapeutic Chemical level 4; **RR** = reporting ratio; **SU** = statistical unexpectedness.

decision rules (table IV), the number of those unmasked SDRs that involved the 29 drugs in the reference event database for the corresponding protocols was 1, 2, 3 and 4, respectively. None of the SDRs gained with empirical protocols involved the 29 drugs with level of evidence information. The level of evidence of masked SDRs by protocol is displayed in table VI.

Discussion

We confirmed previous predictions that masking can occur in any database, even large health authority databases. Our drug selection and removal protocols based on statistical decision rules elicited 38–156 masked SDRs, representing a 0.43–1.8% increase over the 8719 baseline SDRs in AERS involving the 63 selected events. There are few published results against which to compare our findings. In one previous case study involving the event pancreatitis in the US AERS database it was not possible to elicit masked SDRs.[5] Zeinoun et al.[6] found that 0.6% of vaccine-event pairs in a pharmaceutical company vaccine adverse event database demonstrated at least a 2-fold increase in RR after removal of the one vaccine out of five associated with 40% of all reports. The aforementioned percentage gain is within the range we observed for our SU-based protocols. However, the results are not directly comparable. In addition to the obvious differences in the nature of the respective datasets and criteria for report removal, the requirement of Zeinoun et al.[6] for a 2-fold increase may be considered rather severe. We chose to consider masking that induced a change from a subthreshold to a suprathreshold RR since this corresponds

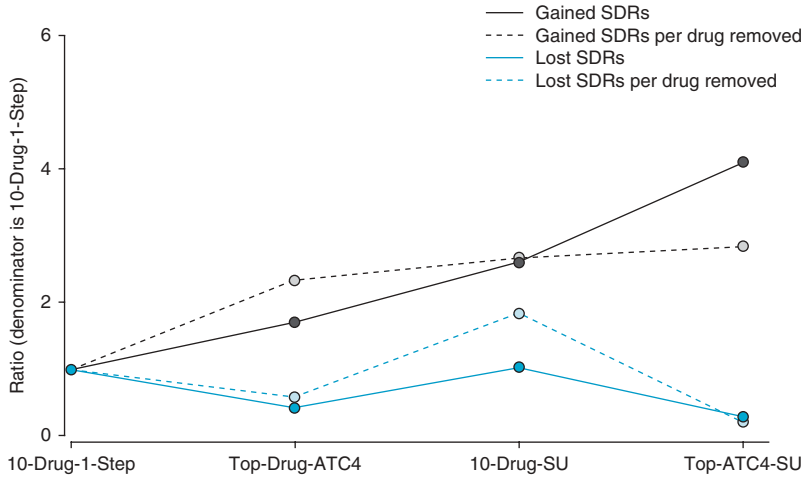

Fig. 2. Per protocol number of gained/lost signals of disproportionate reporting (SDRs) and gained/lost SDRs per median number of drugs removed relative to 10-Drug-1-Step protocol. **ATC4** = Anatomical Therapeutic Chemical level 4; **SU** = statistical unexpectedness.

to the most common signal detection application of disproportionality analysis in pharmacovigilance. Drug-event pairs near threshold would require much less than the 2-fold change in RR required by Zeinoun et al.[6] who may have therefore underestimated the extent of masking from that

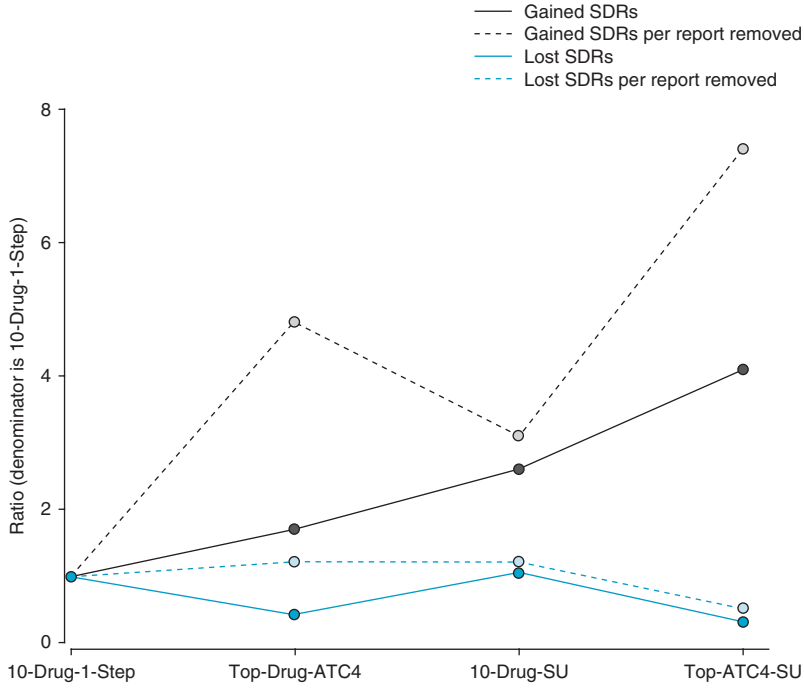

Fig. 3. Per protocol number of gained/lost signals of disproportionate reporting (SDRs) and gained/lost SDRs per median number of reports removed relative to 10-Drug-1-Step protocol. **ATC4** = Anatomical Therapeutic Chemical level 4; **SU** = statistical unexpectedness.

Table V. Gained and lost signals of disproportionate reporting (SDRs) for myoclonus for statistical unexpectedness (SU)-based versus empirical protocols

	10-Step/10-Drug		ATC4	
	SU	empirical	SU	empirical
Gained SDRs	1	10	5	6
Lost SDRs	58	26	5	4

ATC4 = Anatomical Therapeutic Chemical level 4.

perspective. Gould[1] reported that initial RRs for 14 out of 15 reported events that were slightly below threshold exceeded thresholds after the potentially masking drug was removed; however, this was not a database-wide systematic study so we can only guess that this one example suggests a higher incidence of masking in pharmaceutical company databases, as predicted by theory.

The number of masked SDRs generated may be related both to the number and nature of the drugs/reports removed per protocol. To tease apart the effect of number versus nature we normalized the number of gained SDRs to the median number of drugs and reports removed per protocol (i.e. number of gained SDRs per drug removed and per report removed). When normalized this way the difference between protocols based on ATC4 groupings diminished and both showed a proportionately higher number of SDRs generated than the iterative SU-based protocols. Thus, the difference in the number of SDRs elicited by the Top-Drug-ATC4 versus the Top-ATC4-SU protocols was more dependent on the number of reports removed by each than the differences between 10-Drug-SU and 10-Drug-1-Step protocols, which were different from the ATC4-based protocols in both the number and nature of the reports removed.

Compared with a protocol based on the statistical associations of individual drugs, a purely empirical protocol elicited ten times as many SDRs (1 vs 10) for the event myoclonus. If we use this as a very rough guide to the number of masked SDRs in AERs that may have been missed by our SU-based protocols, it is possible that an empirical approach may have elicited approximately 1000 masked SDRs, representing an 11% increase over baseline. However, the computational intensity of this approach precludes its use for agnostic global screening for masking, but it could possibly be used on a limited and *ad hoc* basis in routine pharmacovigilance when the focus of attention is already narrowed to a specific event for example.

Beyond the conventional high-level explanation that larger and more diverse databases such as AERS should be less susceptible to masking, the limited ability of the SU-based protocols to unmask SDRs can be better understood in light of the fact that theoretical formulations of masking in the published literature do not fully incorporate aspects of real-world pharmacovigilance databases and quantitative signal detection procedures. The arithmetic of masking in the published literature does not accommodate the complex interdependencies resulting from multiple drugs and multiple events being recorded in the same report. For example, per protocol removal of a potentially masking drug may simultaneously remove reports of the event under study with the potentially masked drug as well, and removal of reports with one drug can change counts in all four cells of the corresponding 2×2 table. Data on gained versus lost SDRs suggest that these interdependencies may act as an internal damping mechanism to further limit the number of masked SDRs uncovered. We note that those SDRs are

Table VI. Level of evidence for gained signals of disproportionate reporting (SDRs) by protocol

Protocol	Original label	No evidence	Minimal evidence	Possible	Probable	Definite	Total
Top-ATC4-SU	2	1	0	1	0	0	4
10-Drug-SU	1	2	0	0	0	0	3
Top-Drug-ATC4	0	1	0	1	0	0	2
10-Drug-1-Step	1	0	0	0	0	0	1

ATC4 = Anatomical Therapeutic Chemical level 4; **SU** = statistical unexpectedness.

not really 'lost' unless an analyst performed only one analysis for purposes of unmasking. One could evaluate all SDRs from the full database and then evaluate SDRs from versions of the database with certain drugs eliminated. Furthermore, the published formulations of masking do not take into account the typical use of measures of variability (e.g. χ^2 and SU) in parallel with disproportionality measures to define an SDR. Bayesian methods also take account of variability, both global and for the specific DEC under study, though in a different way. Nonetheless, the large number of baseline SDRs lost through a decrease in SU suggests that increased variability may also constrain the number of masked SDRs that can be uncovered by drug removal in AERS. It would be interesting to determine how distributional assumptions and modelling of variability can impact exercises to elicit unmasking.

We did not constrain our concept of masking to related drugs as the concept was originally formulated. However, we did examine this element of masking, most specifically with the Top-Drug-ATC4 protocol. The original literature on masking has implicitly advocated a 'what-if' exercise of eliminating related drugs and regarding the resulting unmasked signals as valid, although relatedness has also been previously based on purely quantitiative considerations. Zeinoun et al.[6] removed vaccine reports based purely on quantitative considerations, which is understandable since relatedness may be a different concept with different significance for vaccines than for drugs. No convention has been established regarding the desirability and validity of signals resulting from eliminating unrelated drugs. None of the gained SDRs involved drugs from the ATC4 grouping of the removed drug. The results do not disprove the relevance to masking of relatedness between drugs and may just reflect the small contribution of any single ATC4 group as a source of masked SDRs.

The criteria of pharmacological relatedness invoked in the 'classical' formulation of masking[1] raises an interesting practical question. Signal detection incorporates multiple lines of evidence from multiple data streams. In addition to quantitative measures, there are multiple other elements considered before nominating an association as a signal. One of these is biological plausibility. Therefore, while discovering a masked SDR on the basis of relatedness to another drug known to cause an event may in a sense be more persuasive (both positive quantitative and biological evidence present), it arguably provides less novel information from an exploratory data perspective since the association involving the potentially masked drug may have been considered a potential signal based on biological plausibility arguments alone. On the other hand, a masked SDR for an unrelated drug that would not otherwise have triggered a suspicion based on biological plausibility considerations, while less persuasive, may arguably provide more interesting exploratory information.

The level of evidence in support of the masked associations was never more than 'possible' using an inclusive level of evidence hierarchy and were often already labelled or had no or minimal supporting evidence. Integrated with the previous findings and their significance as discussed above, it does not appear that routine, global, agnostic searches for masked SDRs in large health authority databases is a productive use of pharmacovigilance resources.

We acknowledge several additional limitations of our study that could have affected the number of masked SDRs elicited and the interpretation of our findings.

We could have elicited additional masked SDRs by removing more drugs but we set reasonable limits to avoid data torturing; however, we also calculated the number of masked associations uncovered *per* drug and *per* report removed. Furthermore, it was not our intent to determine if the protocols that we employed were the 'most effective' of all possible protocols for identifying masking. However, as discussed later in this section, an important result of our investigation is that the current 2×2 table-based formulations of masking do not capture the realities of contemporary pharmacovigilance datasets, where masking effects do not act in isolation and so the concept of an 'optimal' or 'most effective' protocol may not be straightforward in this context.

We used a specific implementation of a single data mining algorithm, and cannot directly make statements about the occurrence of masking with

other algorithms or implementations. For example, there are certainly other statistical measures besides SU that incorporate both disproportionality and case count, such as the 5% lower confidence limit of RR, and top-ranking by case count is another option. Bayesian shrinkage is another common approach to combining disproportionality and case count in a single measure. The investigation of different algorithms and methods for choosing the most influential drug is a potential topic for future research. As discussed in a recent state-of-the-art editorial in *Drug Safety*,[10] different algorithms are currently applied to different databases in pharmacovigilance, and generating convergent findings from multiple study settings may be especially useful by virtue of their generalizability.

However, because of the fundamental principles of disproportionality analysis involved, it is uncertain whether there would be dramatically different results for other algorithms based on how they calculate an observed/expected (O/E) reporting frequency ratio or an RR. Furthermore, we measured gained SDRs relative to baseline so that differences in the propensity to generate SDRs in general between algorithms or specific implementation of a given algorithm might have a comparable effect on both the number of baseline and gained SDRs. However, different data mining algorithms have common implementations involving different SDR-defining thresholds, and the distribution of RRs around these thresholds could produce different results.

An interesting corollary to the issue of algorithm selection is that performing this study would be much more challenging to complete with data mining algorithms such as the multi-item gamma-Poisson shrinker, which are computationally intensive. The practical implications of computational intensity has been reviewed in previous publications but our experience indicates that it has implications for research as well.[11]

Multivariate methods such as shrinkage regression may have advantages in this regard. In fact, it has been used to study the converse 'innocent bystander' phenomena,[12] although our study provides results that might be more representative of the experience of users of 2×2-based approaches in real-world pharmacovigilance scenarios.

We used membership in the ATC4 drug group as a surrogate for relatedness that has not been validated, and ATC3 would have been another rational choice that may have resulted in more unmasking since the latter is higher up in the ATC hierarchy and would constitute larger groups of drugs for removal. Our choice of ATC4 was based on rational considerations and was supported by a limited qualitative review of ATC drug classifications in the WHO dictionary. However, the correlation between the intrinsic characteristics of pharmacophores responsible for adverse reactions and characteristics captured by the ATC classification scheme are unknown and probably variable. In some instances, ATC4 is possibly overly granular, and in others the ATC3 was overly broad.

We are aware that the AERS database and similar databases contain some DECs with very high counts, in some cases as a result of well publicized adverse events and possibly medico-legal considerations.[13] These DECs could be influential in a way that could lead to masking (not to mention false positive findings), but we did not deliberately seek these out. Other anomalies such as 'extreme duplication'[14] may act similarly. So while we reassure the reader that masking is uncommon, we also caution the reader to know the events associated with their drugs under study and to know something about the spectrum of drugs associated with those events – at least enough to identify atypical situations where a secondary analysis for masking may be indicated.

The major references on masking describe the removal of all reports with the potentially masking drug. Another option that, to our knowledge, has not been discussed is the removal of the subset of reports of the specific DEC, which could have a slightly greater effect. In that case one would just identify the drug with the largest number of reports of the DEC of interest.

We used a subset of the FDA AERS database representing a 5-year period. In real-world pharmacovigilance, the entire AERS database, or significantly larger subsets of the data (e.g. post-1997 after the Coding Symbols for Thesaurus of Adverse Reaction Terms [CoSTART]-MedDRA® switch) are typically used. Each might provide

different results for multiple reasons. The results of our study may not necessarily apply to company pharmacovigilance databases where a single drug related to the 'Drug A' under study may be more influential. Performing similar experiments as described in this paper on a company database would be an interesting avenue for future research; however, the limited number of masked SDRs elicited per protocol combined with the pattern of lost SDRs suggests additional common mechanisms could apply to pharmaceutical company databases as well. In fact, Zeinoun et al.[6] reported that 0.4% of vaccine-event pairs decreased by at least 2-fold in their exercise to elicit masking in a company vaccine adverse event database.

The small number of SDRs that involved the 29 reference event drugs was a rate-limiting step in coming to any meaningful conclusions about the level of evidence supporting unmasked associations. Another limitation for this element of the study is that the reference event database used for the level of evidence analysis is not continuously curated and it is possible that some associations have changed evidence categories in either direction.

This study incorporated a subjective step of deciding whether SDRs were due to confounding or bystander effects. Only a small fraction of SDRs were affected by these decisions, but subjective assessments in pharmacovigilance may not be highly reproducible among individual raters.[15]

Interestingly, the more we tried to conceptually grasp and physically excavate some true underlying population of masked SDRs the more elusive it seemed. Different protocols will unmask different SDRs and the true underlying population of masked SDRs would be all those that could be unmasked by all possible unmasking protocols, which is elusive. We do, however, get a sense of what may occur using rational and practically implemented protocols.

We evaluated masking in the context of signal detection. Removing reports could be a potentially useful pharmacovigilance analysis option for signal evaluation. For example, if one were examining the reported association of omeprazole and cough[16] (possible confounding by the indication of gastroesophageal reflux disease aside), for purposes of this discussion, secondary analysis might entail removing reports from pharmacologically/therapeutically disparate drugs heavily associated with cough, such as ACE inhibitors and various inhaled medications. If initial data mining did not generate an SDR, this negative finding might be viewed as more persuasive if it persisted after repeated attempts to elicit masking. It might be possible to derive an 'index of unmasking resistance' to further understand such negative findings.

It is difficult to make a definitive recommendation as to whether unmasking should be routinely performed in pharmacovigilance and, if so, how one should go about it. The phenomenon is briefly mentioned in one regulatory guidance document that does not offer a recommendation.[3] A white paper from an expert working group provides a more extended discussion and chooses to avoid 'blanket recommendations', but suggests checking for masking when suggestive report distributions are present.[2] Our results indicate that considerable resources could be expended on unmasking a small number of SDRs with low positive predictive value for truly novel causal associations. Presently, our conclusion is that routine global agnostic exercises to elicit masked SDRs are unwarranted and that attempts to unmask associations should be knowledge driven.

Conclusions

Attempts to elicit masking using practically implementable protocols produced small changes in the output of SDRs in our analysis of the FDA AERs database. This is undoubtedly related to the large size and diversity of the database, but the complex interdependencies between drugs and events in authentic spontaneous reporting system (SRS) databases and the impact of commonly used measures of statistical variability may be additional factors constraining the discovery of masked SDRs that may also operate in any SRS database. Importantly, our investigation demonstrated that the complex interdependencies in pharmacovigilance databases that are not accommodated in current formulations of masking and

2×2 table-based methods have significant impacts on masking. Therefore, multivariate methods, such as shrinkage regression, may have more utility in this setting.

Routine agnostic exercises to elicit masking using 2×2 table-based methods, especially in large health authority databases, is not justified based on results available to date. Exercises to elicit unmasking should be driven by prior knowledge or obvious data imbalances. When practical, analysts should know the events associated with their drugs under study and know something about the spectrum of drugs associated with those events – at least enough to identify atypical situations where a secondary analysis for masking may be beneficial.

Extension of the definition of classical masking to one based purely on numerical reporting relationships that transcends the requirement for pharmacological/therapeutic relatedness between the masked and masking drug is intuitively plausible and may ultimately prove to be more informative from the perspective of exploratory data analysis.

Acknowledgements

We thank the anonymous reviewers of this manuscript for their many helpful comments.

This work was partly funded by a grant from the Pharmaceutical Research and Manufacturers Association (PhRMA) to ProSanos Corporation. At the time of writing this manuscript, Alan Hochberg and Ronald Pearson were employed by ProSanos Corporation and, as employees of ProSanos, have patents pending on data mining visualization tools, assigned to ProSanos Corporation. Alan Hochberg is presently employed by F. Hoffman-La Roche Ltd, Basel, Switzerland, and Ronald Pearson is presently employed by The Travelers Company, Inc., Hartford, CT, USA. Manfred Hauben and Hsin-wei Wang are full-time employees of Pfizer, Inc., which manufactures and/or markets drugs discussed in this article and/or potentially competing drugs in the same therapeutic classes as drugs discussed in this article. They own stock and stock options in Pfizer Inc. and own stock in other pharmaceutical companies for which the above-described competing interests may apply.

Manfred Hauben, as a representative of the funding committee of PhRMA, participated in the design of the study, the interpretation of data and the generation and review of the manuscript. Hsin-wei Wang, Alan Hochberg and Ronald Pearson were responsible for the design and conduct of the study, data collection, management and analysis, interpretation of the data and generation of the manuscript.

Appendix

Description of Masking

To illustrate the basis of masking, consider the contingency table below (table AI). From this table, we can compute the RR as a measure of disproportionality. Other data mining methods use different measures, but RR will serve here for illustrative purposes.

Table AI. Contingency table for disproportionality calculation

Drugs listed per report	No. of reports		
	event mentioned	event not mentioned	total
Drug A mentioned	500	20 000	20 500
All other reports	12 500	1 020 000	1 032 500
Total	**13 000**	**1 040 000**	**1 053 000**

For this table, the RR for Drug A and the event of interest is calculated as $RR = (500/20\,500) \div (13\,000/1\,053\,000) = 1.98$. In this case, using the criteria that an RR >2.0 indicates an 'SDR', i.e. an indication of a statistical association between Drug A and the adverse event under study, we would say that there is no SDR for Drug A and the event under study.

Now consider the table below (table AII), which differs from the one above only in that we have separately listed reports for a particular drug, Drug B, for which the adverse event under study occurs frequently.

Table AII. Contingency table for illustrating masking

Drugs listed per report	No. of reports		
	event mentioned	event not mentioned	total
Drug A mentioned	500	20 000	20 500
Drug B mentioned	2 500	20 000	22 500
All other reports	10 000	1 000 000	1 010 000
Total	**13 000**	**1 040 000**	**1 053 000**

For simplicity in this example, let us assume that mention of Drug A and Drug B are mutually exclusive. If we were to eliminate reports for Drug B, the RR for Drug A and the event of interest would be $RR = (500/20\,500) \div (10\,500/1\,030\,500) = 2.39$. This is above the commonly

used detection threshold of RR ≥2.0, and therefore we would declare that there is an SDR for Drug A and the event under study. The SDR was 'masked' by the presence of Drug B. The reason for the masking is the large number of reports for Drug B that mention the event under consideration. This number is large enough to significantly influence the fraction of reports in the entire database that mention the event.

References

1. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf 2003; 12: 559-74
2. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf 2005; 28 (11): 981-1007
3. EMEA Eudravigilance Expert Working Group (EV-EWG). Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system. 26 June 2008. London. Doc. Ref. EMEA/106464/2006 rev.1 [online]. Available from URL: http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2009/11/WC500011434.pdf [Accessed 2010 Sep 6]
4. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005 Sep; 4 (5): 929-48
5. Hauben M, Hochberg A. The importance of reporting negative findings in data mining: the example of exenatide and pancreatitis. Pharmaceut Med 2008; 22 (4): 215-9
6. 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 (9): 599-607
7. Hauben M, Aronson JK. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. Drug Saf 2009; 32 (2): 99-110
8. Hauben M, Reich L. Response to letter by Levine et al. Br J Clin Pharmacol 2005; 61 (1): 115-7
9. Hochberg AM, Reisinger SJ, Pearson RK, et al. Using data mining to predict safety actions from FDA adverse event reporting system data. Drug Inf J 2007; 41 (5): 633-44
10. Hauben M, Noren GN. A decade of data mining and still counting. Drug Saf 2010; 33 (7): 527-34
11. Hauben M, Madigan D, Hochberg AM, et al. Data mining in pharmacovigilance: computational cost as a neglected performance parameter. Int J Pharm Med 2007; 21 (5): 319-23
12. Caster O, Noren GK, Madigan D, et al. Large scale regression-based pattern discovery: the example of screening the WHO global drug safety data base. Stat Anal Data Min 2010; 3 (4): 197-208
13. Pariente A, Gregoire F, Fourrier-Reglat A, et al. Impact of safety alerts on measures of disproportionality in spontaneous reporting databases: the notoriety bias. Drug Saf 2007; 30 (10): 891-8
14. Hauben M, Reich L, DeMicco J, et al. 'Extreme duplication' in the US FDA Adverse Events Reporting System Database. Drug Saf 2007; 30 (6): 551-4
15. Arimone Y, Miremont-Salamé G, Haramburu F, et al. Inter-expert agreement of seven criteria in causality assessment of adverse drug reactions. Br J Clin Pharmacol 2007 Oct; 64 (4): 482-8
16. Howalzi M, Delafosse C. Omeprazole-induced intractable cough. Ann Pharmacother 2003; 37: 1607-9

Correspondence: Dr *Manfred Hauben*, Pfizer Inc., 150 East 42nd Street, New York, NY 10021, USA.
E-mail: Manfred.Hauben@Pfizer.com