Drug Safety 2007; 30 (6): 551-554
0114-5916/07/0006-0551/$44.95/0

'Extreme Duplication' in the US FDA Adverse Events Reporting System Database

We recently encountered an example of extreme duplication in the publicly released version of the US FDA Adverse Events Reporting System (AERS) database, which is available through the Freedom of Information (FOI) Act. We use the term 'extreme duplication' because the majority of reports of the drug-event pair under study were found to be duplicate reports, and these duplicates resulted in the occurrence of a large signal of disproportionate reporting (SDR)[1] discovered during a data-mining exercise. This occurred during a recent demonstration within our company of a commercial vendor's data-mining software containing these data. Data mining includes emerging computer-based quantitative tools for screening spontaneous reporting system (SRS) databases to assist in the identification of potential new drug safety issues. Although duplicate reporting is one of the numerous well recognised forms of data corruption and distortion in SRS databases, the finding surprised us nonetheless, given the duplicate detection algorithms and procedures included in the commercial data-mining software we used.

In the demonstration, one drug was randomly selected as a drug of interest (i.e. cases reporting the selected drug as suspect drug through the fourth quarter of 2005) and all adverse event terms for that drug were reviewed using a configuration where duplicate reports were removed by the vendor via the aforementioned duplicate-detection procedures. Data-mining results indicated an especially strong SDR for the event 'aortic dissection', prompting us to further consider this finding. For the year 2005, there were 66 cases in total reporting aortic dissection, of which 20 were reported with the drug of interest. All 20 cases reported the same event date, the same 11 suspect medications, the same outcome and the same gender. In all 20 cases, 'age' was not reported. After review of the case detail fields in the vendor's programme, we concluded that the cases were all duplicates of the same incident and the finding was an example of extreme duplication. Upon review of a representative case narrative, there was no substantive evidence of a causal association with aortic dissection and any of the 11 suspect drugs.

We ran the same query with a similar configuration using a second vendor's software (for the second vendor, the user has several options for removing duplicate cases; we chose the most aggressive option) and came up with the same 20 cases. However, when a similar query (there is no event term aortic dissection in the WHO dictionary so the default WHO term 'aneurysm' was used instead) was run in Vigibase (the database of the WHO Programme for International Drug Monitoring), which is a multinational safety database that receives reports from regulatory authorities around the world including the US, none of the 20 relevant reports were identified. We suspected that, at the time of our demonstration, these reports had not as yet been transmitted to WHO by the FDA.

The major reasons that report duplication can occur in regulatory and/or company databases are: (i) reporting from different sources (e.g. health professionals, pharmaceutical companies and consumers having provided separate case reports related to the same incident) and failure to link these cases; and (ii) failure to link up follow-up case reports to earlier records.

We believe that two circumstances most probably accounted for our observed extreme duplication. First, this incident was received by one pharmaceutical company and forwarded to other pharmaceutical companies that manufactured the co-suspect and concomitant medications or was received by a regulatory agency from a manufacturer and then forwarded by the regulatory agency to other manufacturers.

In the US, healthcare professionals and consumers can voluntarily report adverse events to the FDA. Adverse events are also reported to pharmaceutical manufacturers; however, the manufacturer has a statutory obligation to forward the adverse event to

the FDA. Pharmaceutical manufacturers can receive reports where a patient is taking multiple suspect medications marketed by different manufacturers. Under FDA guidance (i.e. Guidance for Industry), they have been asked to forward the adverse event to the other manufacturers;[2] all manufacturers will then have to report the adverse event to the FDA.

The AERS data are purchased by vendors, downloaded into their software and then supplied to their customers in quarterly updates. The vendors, using a combination of manual inspection and proprietary automated techniques, make significant efforts to pre-process the data to remove/identify duplicate cases before providing to their customers. However, in reports where key criteria are not reported, vendors' programmes may not identify/remove the duplicate cases. The criteria in one vendor's programme that are used for identifying and removing duplicate cases involves querying key fields, such as age, event date, gender and manufacturer control number.[3] In the AERS database in this instance, the 20 cases were reported by 10 different manufacturers with 1–3 reports per manufacturer, apparently indicating duplication across and within manufacturers. Secondly, the patient's age in this particular case was not recorded. The criteria that are applied for removing duplicate cases did not appear to be sufficient when a key field, i.e. age, was not reported for the same event reported by multiple manufacturers in this instance.

The exact extent of duplication, extreme or not, present in AERS and other agency databases is unknown, and there is a paucity of published data examining the issue. In an evaluation of quinine-induced thrombocytopenia, FDA researchers identified 20% of 141 reports as duplicates.[4] To the best of our knowledge, there are no published data on the impact of report duplication on generation of SDRs, although Norén et al.[4] opine that since commonly used data-mining procedures may highlight associations with as few as three reports, one or two duplicates may severely affect their utility. Given the mechanistic link between the recorded suspect polypharmacy and the report duplication in our case and the high prevalence of polypharmacy in general, it is possible that this is not an isolated phenomenon. In fact, a recent analysis of an extract of AERS data from 2001–6 revealed the presence of 103 856 individual safety reports (ISRs) listing at least three suspect drugs. Given the contribution of both multiple suspect drugs and unrecorded age to our example of extreme duplication, it is also interesting to note that the same analysis revealed 28 708 ISRs involving three or more suspect drugs plus unrecorded age (Pearson RK, ProSanos Corporation, personal communication). Parenthetically, while writing this paper, we serendipitously discovered an additional example of apparent extreme duplication, in which five of six reports of toxic epidermal necrolysis appeared to represent the same case, although the explanation for the redundant reporting is unclear, emphasising multiple processes can result in report duplication. Another notable feature of the latter duplication was that the time span between the initial report and latest duplicate was 10 years.

The impact on the data-mining findings in our example of extreme duplication in AERS compared with WHO can be seen in table I, where we provide the results of an analysis of data that were mined using the second vendor's software using the latest data available to us (AERS, second quarter of 2006 and WHO forth quarter of 2006). WHO data were included in our analysis because, as previously mentioned, when we ran our initial query in the WHO database (June 2006), we were unable to find any of the 20 relevant cases. However, in December of 2006, we queried the WHO database again and were able to then find four cases originating from the US with the same 11 suspect medications.

As can be seen from table I, the extreme duplication in AERS lead to SDRs with all 11 co-suspect drugs, severely impacting the data-mining results. When duplicates were excluded from the analysis, there were no SDRs. For WHO, duplicate cases did not have an impact on SDR generation.

Attempts are being made to enhance duplicate detection.[4] For example, a 'hit-miss' model based on a duplicate detection algorithm has been implemented by WHO to assist in the identification and removal of confirmed duplicates from their

Table I. 'Extreme duplication' data-mining findings using data from the US FDA Adverse Events Reporting System (AERS) database[a] and the WHO database[b] for aortic dissection/aneurysm[c] and 11 co-suspect drugs[d]

	AERS data with all cases (duplicates included)	AERS data with duplicates removed[e]	WHO data with all cases (duplicates included)	WHO data with duplicates removed[f]
Mean (± standard deviation) observed counts, 11 co-suspect drugs (range)	19.8 ± 1.9 (16–23)	2.0 ± 1.8 (1–7)	9.6 ± 9.6 (4–33)	6.6 ± 9.6 (1–30)
Mean (± standard deviation) scores of EB05[g] all 11 co-suspect drugs (range)	**16.7** ± 12.8 (**3.8–46.2**)	0.3 ± 0.2 (0.1–0.8)	0.6 ± 0.3 (0.3–1.2)	0.3 ± 0.3 (0.05–1.07)

a Up to the second quarter of 2006.

b Up to the fourth quarter of 2006.

c The term 'aneurysm', the default term for 'aortic dissection' was used in the WHO database.

d The co-suspect drugs were digoxin, paracetamol (acetaminophen), allopurinol, pantoprazole, tiotropium, simvastatin, furosemide, perindopril, warfarin, salbutamol (albuterol) and beclometasone.

e One case number common to all co-suspect drug cases retained; 22 duplicate case numbers excluded.

f One case number common to all co-suspect drug cases retained; three duplicate case numbers excluded.

g EB05 was representative data-mining metric chosen. An EB05 >2 (values highlighted in bold) is a commonly cited threshold considered to be a signal of disproportionate reporting.

EB05 = empirical Bayesian metric.

database.[4] Based on the mechanism of duplication in this instance and the design of the hit-miss model, in our example, the WHO approach might have been effective in detecting duplicates in this type of scenario, although this remains unconfirmed.

Public safety data have been collected over many years from many different sources that do not represent a random sampling mechanism. As a result, they contain a number of inconsistencies, errors, duplications and lack of data in key fields. Despite these limitations, they remain a cornerstone of postmarketing surveillance activity and are used as data sources in research publications. However, these limitations, which are generally acknowledged in the publications, can degrade the capacity for optimal data mining and analysis. Allowing duplicate reports to appear in query results may be misleading but potentially easy to identify. What may be more illusive is the number of credible associations that may be quantitatively masked due to inflation of expected counts by duplicate reporting.

It should be noted that although conventional drug safety surveillance and data-mining approaches, based on disproportionality screens, consider duplicates as noise that should be removed if possible, a recent suggestion that numbers and types of duplicates in spontaneous reporting systems might have independent information value may warrant empirical exploration (Wise RP, US FDA at the 22nd International Conference on Pharmacoepidemiology and Therapeutic Risk Management, August 24-27, 2006, Lisbon, Portugal, personal communication).

Extreme anomalies naturally capture our attention; however, we stress that inferences should not be drawn by an isolated example no matter how interesting or novel – indeed the novelty may suggest that it is an oddity of little practical significance in general. However, it may legitimately serve to raise questions that facilitate knowledge discovery and process improvements. Some experienced investigators have claimed that duplicate reports do not have a significant impact on data mining with spontaneous reports. This may be true, but given the quantitative and qualitative gaps in SRS data, it is also possible that the observed failure of the most commonly used commercial duplication detection procedures is not an isolated anomaly. The Pharmaceutical Research and Manufacturers of America (PhRMA) has funded two major research proposals (principal investigators Alan Hochberg, ProSanos Corporation; Sheila Weiss, University of Maryland

School of Pharmacy/Victor Gogolak, DrugLogic Inc.) designed to understand both the strengths and limitations of data mining in pharmacovigilance including the potential impact of deficits in data quality. Based on our anecdotal experience, the practical impact is difficult to quantify. In this instance, an inquiring mind 'drilling down' logically and empirically identified the issue fairly easily though confirmation was still required expending time. However, there may be different forms of case duplication, some less obvious. Therefore, the deleterious impact on process flow will depend on the number of times such situations occur and the ease with which the duplication is confirmed. Having a full awareness of the limitations of the data, and the inability of statistical modelling and current proprietary de-duplication procedures to neutralise these limitations, it is more important than ever, given the precedent for reporting of data-mining findings in the scientific literature, popular press and the courtroom, based on aggregate SRS data without case-level clinical review.[5,6]

Given the potential for commercial and ideological conflicts of interest associated with data mining in pharmacovigilance, and the tendency for over-enthusiasm that technology may generate, it is not surprising that Bate and Edwards[7] identified persistent views about data mining in pharmacovigilance that include extremes of 'unbridled optimism' and 'considerable scepticism'. In addition to identifying a potential area for reporting process improvement and a potentially useful topic for further investigation, our objective in presenting this example of 'extreme' duplication, is to add further support to a perspective that is intermediate between the above extremes. We strive to fully understand and acknowledge both the strengths and limitations of the data and the methods, and emphasise the need to consider any findings based on SRS data with an appropriate level of caution. Promoting or using data-mining software as an alternative to clinical comprehension can have deleterious consequences for patient safety.[6]

Manfred Hauben,[1,2,3,4] *Lester Reich*,[1] *James DeMicco*[1] and *Katherine Kim*[1]

1 Safety and Risk Management, Pfizer Inc, New York, New York, USA
2 Department of Medicine, New York University School of Medicine, New York, New York, USA
3 Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, Valhalla, New York, USA
4 School of Information, Computing, and Mathematics, Brunel University, West London, UK

Acknowledgements

Pfizer markets/co-markets one or more of the drugs cited in this letter.

References

1. Hauben M, Reich L. Communication of findings in pharmacovigilance: use of the term "signal" and the need for precision in its use. Eur J Clin Pharmacol 2005; 61: 479-80
2. CDER Guidance. Guidance for industry: guideline for postmarketing reporting of adverse drug experiences, 85D-0249 [online]. Available at URL: http://www.fda.gov/medwatch/REPORT/Mfg.htm [Accessed 2006 Dec 28]
3. Preparing FOI Drug Safety Data for WebVDME Data Mining, Lincoln Technologies Inc., 2005
4. Norén GN, Orre R, Bate A, et al. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Disc 2007; 14 (3): 305-28
5. Strom BL. Evaluation of suspected adverse drug reactions. JAMA 2005; 293 (11): 1324-5
6. Hauben M, Patadia V, Gerrits CM, et al. Data mining in pharmacovigilance: the need for a balanced perspective. Drug Saf 2005; 28 (10): 835-42
7. Bate A, Edwards IR. Data mining in spontaneous reports. Basic Clin Pharmacol Toxicol 2006; 98 (3): 324-30

Correspondence: Dr *Lester Reich*, Risk Management Strategy, Pfizer Inc, 235 E. 42nd Street, New York, NY 10017, USA.
E-mail: Lester.Reich@Pfizer.com