Letter to the Editor

Response to letter by Levine *et al.*

Manfred Hauben^{1,2,3} and Lester Reich¹

¹Risk Management Strategy, Pfizer Inc, New York, USA;

²Department of Medicine, New York University School of Medicine, USA; ³Departments of Pharmacology and Community and Preventive Medicine, New York Medical College, USA

Levine, Tønning, and Szarfman provide a critique [1] of our publications in which we shared observations and concluding discussions consistent with those of other researchers (C. Gerrits, personal communication) [2–7]. Some of the richest learning experiences arise from the free and lively exchange of ideas expressed in letters to the editor, so we feel privileged to participate in a dialogue that may help illuminate the current state of data mining in pharmacovigilance (PV). Herein, we address specific statements and posthoc analysis/interpretations in their letter. To comply with editorial limitations, we can only address key points.

Levine *et al.* misconstrued our investigations as being based on preset assumptions about data mining performance. Our letters identified adverse events (AEs) that are of demonstrated interest in PV and signal detection because of their clinical significance and level of documentation. We addressed the relevant question of whether one safety reviewer using a data mining procedure based on proportional reporting ratios (PRRs) and one using a procedure based on the multi-item gamma Poisson shrinker (MGPS) could have been alerted to the AEs concurrently or at different time points. The data revealed that with commonly used and recommended levels of disproportionality (including those recommended by Szarfman [8]), a safety reviewer using the procedure based on PRRs could have been alerted to some events in advance of a safety reviewer using the procedure based on MGPS, and in some instances only the reviewer using the PRRs would have been alerted; these findings are consistent with those of other researchers from diverse settings who have studied the trade-offs in sensitivity and specificity between PRRs

and the (multi-item) gamma Poisson shrinker, (M)GPS, using commonly cited thresholds (C. Gerrits, personal communication) [2–6]. An oft-cited characteristic of MGPS is to statistically down-weight associations with low report counts. We discussed the importance of both sensitivity and specificity, that the enhanced sensitivity of PRRs in these settings is associated with reduced specificity that may require additional triage criteria for practical implementation, cautioned against generalizing from our preliminary observations, and recommended further research studying numerous thresholds in diverse PV scenarios.

Their letter questioned our selection of adverse events for measuring data mining performance. They considered trimethoprim/sulphamethoxazole (TMP/SMX)-associated hyperkalaemia unsuitable because a ‘strong signal’ for this AE was due to administration of high-dose TMP/SMX therapy in immunocompromised patients that postdated the early PRR elevations. This illustrates the hazards of evaluating data mining results without a thorough examination of the published literature, which describes the frequent occurrence of TMP/SMX-associated hyperkalaemia in *healthy* outpatients treated for urinary tract infection and hospitalized *non-immunocompromised* patients treated with standard dose therapy for varied infections [9–13]. Further, the potential of an algorithm to highlight relevant associations within spontaneous reporting systems (SRs) prior to widespread recognition in the published literature is an important consideration given the potential for under-reporting to SRs, due in part to under-recognition by potential reporters. Their suggestion that this event was studied because it was added to the TMP/SMX label is incorrect. Our stated selection criterion was persuasive clinical and molecular pharmacological evidence favouring causality. As another example of how ‘*the authors assume that the drug-event combinations (DECs) chosen for analysis represent valid, true-positive drug–event associations*’ when ‘*there is no guarantee that the DECs chosen by the authors are in fact true positives (causally related)*’ their letter

describes our selection criteria for drug-associated pancreatitis as follows: *‘the authors select the association of pancreatitis with various drugs based on upon “definite causal relationships” from external published reports and observational studies . . .’*. This seems misleading; therefore we reproduce the pertinent statement from the relevant letter:

‘Many drugs have been associated with pancreatitis. Published reports and observational studies have asserted definite causal relationships with various drugs including [list of medications]. Although residual uncertainty about causality may still remain even in well documented case reports involving positive rechallenge, it provides an interesting and diverse data set for examining the performance of data mining algorithms (DMAs) since the evidentiary support for the reported associations may be quite strong.’ [14]

The absence of gold standards for adjudicating causality is a crucial issue in data mining and PV, but practical decisions must often be made with less-than-absolute certainty. We maintain that an algorithm can be useful if it helps to identify medically important associations that are *probably* causal, yet not established with meta-physical certitude. Of particular relevance to this point is that the authors used posthoc analyses to claim that MGPS resulted in the same findings uncovered with PRRs regardless of causality.

Their letter took issue with the comparison of discrete disproportionality metrics and inferred that both algorithms provided the same results using a series of unvalidated comparisons of confidence and posterior intervals that have no objective interpretation in this context, and do not address the question that inspired our investigation: whether the time to ‘signal’ would vary between PRR and MGPS when commonly used thresholds of disproportionality were applied. As noted above, comparisons of the same discrete disproportionality metrics are commonly used by researchers in various settings, who have studied the balance between sensitivity and specificity as a function of the thresholds/metrics used (C. Gerrits, personal communication) [2–7]. Although Levine *et al.*’s analyses are not applicable to the practical PV scenarios we studied, they raise the issue of confirmation bias in data mining. For analysts with a strong incentive to expect a particular outcome, the large space of available choices in data mining provides numerous opportunities to choose a metric or procedure based on the fit of the results to their preexisting expectations, in effect retrofitting the analysis to their beliefs. We, as do other investigators (C. Gerrits, personal communication) [2–7] specify in advance the disproportion-

ality thresholds to be studied. Nevertheless, it is gratifying if our work has increased awareness of the subjective and ad hoc nature of these thresholds, and we recommended that methodology be defined in advance.

Levine *et al.* criticized comparing stratified MGPS with unstratified PRRs but carried forward analogous posthoc comparisons using derived confidence intervals and concluded that the stratified MGPS yielded the same results as the unstratified PRR. Because the design of (M)GPS includes a stratified independence model, comparisons between unstratified PRR and stratified (M)GPS is a standard practice used by multiple investigators, who report trade-offs in sensitivity and specificity when commonly cited thresholds are used [2–4]. It is important to note that we have previously considered the relative contributions of Bayesian shrinkage to the null, covariate stratification, and additional statistical manipulations to the observed performance differentials [15]. Although Szarfman has oftentimes emphasized the importance of stratification with MGPS (but has also used unstratified MGPS to support post hoc interpretations [16]), our experience and that of others suggests to us that the significant performance gradients cannot be explained away by basic covariate stratification [2].

Our publications examining sets of multiple drug–event combinations consistently explored disproportionalities obtained with each algorithm across the entire reporting experience. Their letter suggested that most disproportionate PRRs based on one or two reports should be discarded. While we have previously discussed the implications of the sparsity of the US Food and Drug Administration Adverse Event Reporting System (AERS) database and appreciate the precautions needed for interpreting such data, we are exploring a more flexible approach since we have observed that many of these disproportionalities are of much less/more interest when viewed in a clinical context and that synchronous disproportionalities involving clinically equivalent/overlapping AE codes and/or drug formulations in separate, individual reports may be of interest when using PRRs (an important consideration for AERS data, which are coded using the hyper-granular Medical Dictionary for Regulatory Affairs [MedDRA] dictionary). It is also important to be aware of the problem of underreporting in PV. We are interested in exploring the differential impact on PRRs and MGPS of the massive underreporting in SRSs (a phenomenon pertinent to the rhabdomyolysis letter, where despite numerous cases in the index literature report, there was only one in the AERS dataset). Finally, our observations are consistent with those of other researchers, who report trade-offs in sensitivity and specificity between PRRs and MGPS

even when the PRR/χ^2 thresholds is combined with an additional threshold based on case counts (C. Gerrits, personal communication) [2–6]. The assertion that our methodology included relaxing criteria for the level of disproportionality and statistical significance is factually incorrect, and we refer readers to the referenced articles to make their own judgements about this and our other points.

Despite the obvious uncertainty surrounding the interpretation of data mining results, we hope readers will maintain a positive view about data mining. We have no vested interest in any one approach and regard all the contemporary quantitative approaches to knowledge discovery in PV as credible and promising options, including those not discussed here [17–19]. Each method has strengths and limitations [7]. As we stated in our publications, while it is important to understand performance gradients between approaches and their determinants, the practical significance of these differentials is unclear, since thresholds are currently subjective, and should only be used as one element in a comprehensive signal detection programme. We also believe the importance and impact of the trade-off between sensitivity and specificity may vary across PV organizations since they are not structurally and functionally homogeneous. As a result the incremental utility of each algorithm may well depend on the precise mode of deployment as well as the resources, experience, and clinical expertise within the organization.

References

- 1 Levine JG, Tonning JM, Szarfman A. Reply: The evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned. *Br J Clin Pharmacol* 2005; in press.
- 2 Kubota K, Koide D, Toshiki H. Comparison of data mining methodologies using Japanese spontaneous reports. *Pharmacoepidemiol Drug Saf* 2004; 13: 387–94.
- 3 Kawabe E, Sato T, Kubota K. Comparison of methodologies for signal detection using Japanese spontaneous reports. Abstract: 19th ICPE. *Pharmacoepidemiol Drug Saf* 2003; 12 (S1): S169.
- 4 Moseley JNS, Heeley E, Ekins-Daukes S, Evans S. Preliminary comparison of two signal detection methodologies in the UK regulatory spontaneous ADR data base. *Drug Saf* 2004; 27 (12): 950–1.
- 5 Roux E, Tubert-Bitter P, Thessard F, Fourrier A, Begaud A, Evans S. Automatic signal detection methods evaluation on a simulated data base. Workshop: 20th ICPE, 2004.
- 6 Banks D, Woo EJ, Burwen DR, Perucci P, Braun MM, Ball R. Comparison of four data mining methods in the US vaccine adverse event reporting system (VAERS). *Pharmacoepidemiol Drug Saf* 2003; 12 (S1): S138.
- 7 Banks D, Woo EJ, Burwen DR, Perucci P, Braun MM, Ball R. Comparing data mining methods on the VAERS database. *Pharmacoepidemiol Drug Saf* 2005; 14: 601–609.
- 8 Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Saf* 2002; 25 (6): 386–387.
- 9 Allapan R, Perazella MA, Buller GK. Hyperkalemia in hospitalized patients treated with trimethoprim sulfamethoxazole. *Ann Intern Med* 1996; 124: 316–20.
- 10 Allapan R, Buller GK, Perazella MA. Trimethoprim-sulfamethoxazole therapy in outpatients: is hyperkalemia a significant problem? *Am J Nephrol* 1999; 19 (3): 389–94.
- 11 Modest GA, Price B, Mascoli N. Hyperkalemia in elderly patients receiving standard doses of trimethoprim-sulfamethoxazole. *Ann Int Med* 1994; 120 (5): 437.
- 12 Pennypacker LC, Mintzer J, Pitner J. Hyperkalemia in elderly patients receiving standard doses of trimethoprim-sulfamethoxazole. *Ann Int Med* 1994; 120 (5): 438.
- 13 Canaday DH, Johnson JR. Hyperkalemia in elderly patients receiving standard doses of trimethoprim-sulfamethoxazole. *Ann Int Med* 1994; 120 (5): 437–8.
- 14 Hauben M, Reich L. Drug-induced pancreatitis: lessons in data mining. *Br J Clin Pharmacol* 2004; 58: 560–2.
- 15 Hauben M, Reich L. Safety related drug-labelling changes: findings from two data mining algorithms. *Drug Saf* 2004; 27: 735–44.
- 16 Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: New systematic tools for an old problem. *Pharmacotherapy* 2004; 24 (9): 1102.
- 17 Bate A, Lindquist M, Edwards IR, Orre R. A data mining method for signal detection and analysis. *Drug Safety* 2002; 25 (6): 393–7.
- 18 van Puijenbroek EP, Bate A, Leufkens HG, Lindquist M, Orre R, Egberts AC. A comparison of methods of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Safety* 2002; 11 (1): 3–10.
- 19 van Puijenbroek E, Diemont W, van Grootheest K. Application of quantitative signal detection in the Dutch spontaneous reporting system for adverse drug reactions. *Drug Safety* 2003; 26 (5): 293–301.

Received

June 22 2005

Accepted

July 17 2005

Correspondence

Lester Reich, MD, Risk Management Strategy, Pfizer Inc., New York, NY 10017, USA; Email: lester.reich@pfizer.com