Large-Scale Regression-Based Pattern Discovery in International Adverse Drug Reaction Surveillance

Ola Caster
WHO Collaborating Centre for
International Drug Monitoring
Uppsala, Sweden
ola.caster@who-umc.org

G. Niklas Norén
WHO Collaborating Centre for
International Drug Monitoring
Uppsala, Sweden
niklas.noren@who-umc.org

David Madigan
Department of Statistics
Columbia University
New York, USA
madigan@stat.columbia.edu

Andrew Bate
WHO Collaborating Centre for
International Drug Monitoring
Uppsala, Sweden
andrew.bate@who-umc.org

ABSTRACT

This paper demonstrates the first use of shrinkage logistic regression as a pattern discovery method in international adverse drug reaction surveillance. This novel method is compared to bivariate pattern discovery, the standard approach in the application domain. Our results show that regression can eliminate false positives and false negatives due to the impact of other covariates, and that it can retrospectively detect established drug safety issues earlier than bivariate pattern discovery. However, regression cannot completely replace bivariate methods, for two reasons: its failure to identify some established drug safety concerns as early as the routinely used bivariate measures, and the relative intransparency of the procedure to estimate the regression coefficients. This suggests both approaches be used in parallel.

1. INTRODUCTION

1.1 Adverse Drug Reaction Surveillance

Pattern discovery, the identification of interesting patterns in often massive data sets, is a key component of international adverse drug reaction (ADR) surveillance [1, 2, 9]. The cornerstone of ADR surveillance is the collection and analysis of voluntarily submitted reports of suspected ADR incidents in regular clinical practice [1]. The reports are screened for previously unknown possible causal relationships that are clinically judged worthy of further follow-up [1]. We consider the world's largest database of its kind: the 4 million record WHO Global Individual Case Safety Report database, Vigibase, which contains data on about 14,000 unique drug terms and 2,000 unique ADR terms. With the yearly reporting to Vigibase being in the order of 200,000, quantitative screening for excessive reporting is essential in the highlighting of drug–ADR pairs for further clinical assessment [1]. As ADR reports are contributed on a voluntary basis there is not information on all patients who experienced an ADR, nor who received the drug but did not experience a suspected ADR. These limitations in addition to data quality issues means that summary statistics must be interpreted with caution, and collections of ADR reports are not an appropriate basis for hypothesis testing but primarily useful for hypothesis generation.

1.2 Routine and Proposed Methods

In this paper we perform pattern discovery in Vigibase using two methods. As a reference we use the standard method to screen the database, the bivariate Information Component (IC) [1]. In short, the IC is a shrunk lift factor on a $\log_2$ scale, contrasting for each drug–ADR pair the relative reporting rate of the ADR given the drug to the general relative reporting rate of the ADR in the database. For each IC value we calculate a 95 % credibility interval and quote its lower limit, IC_{025} [8]. In routine use, a drug–ADR pair is highlighted for clinical review if $IC_{025} > 0$ [8].

As an alternative method we consider the use of shrinkage logistic regression as exemplified by the open source BBR (Bayesian Binary Regression) software[1] developed by one of the authors [4]. Recent development in the area is not limited to this algorithm alone (see e.g. [5, 6, 11]). Multivariate approaches such as this have previously not been applied to large-scale pattern discovery in ADR surveillance, due to lack of convergence, prohibitive computational complexity, and unstable parameter estimates.

We fit one model for each ADR reported to the database, using each report as an observation. We use as the binary dependent variable the presence of the ADR on the reports, and as binary explanatory variables the presence on the reports of each drug reported to the database. Further, we shrink the parameter estimates by constraining the sum of the absolute values of the β parameters (excluding β_0). We shall refer to this method as Lasso Logistic Regression (LLR), following the special type of so called lasso shrinkage that arises. This shrinkage is characterized by the property that some of the β's are shrunk to exactly zero, whereas the parameters corresponding to strong predictors retain estimates close to their maximum likelihood estimates [12]. This is a very useful characteristic in large-scale pattern discovery, where the computational cost of constructing uncertainty intervals around each β is often prohibitive. The shrinkage to exactly 0 allows us to construct a decision rule for highlighting drug–ADR pairs with LLR based on whether or not the corresponding β estimate is positive. A parameter estimate

---

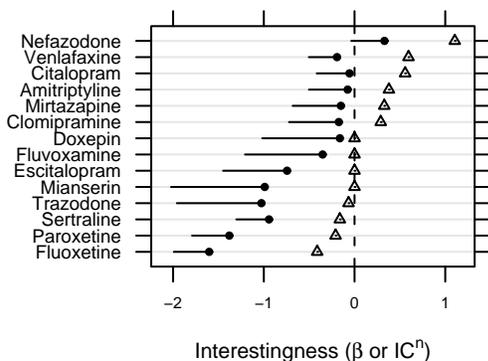[1]Available from www.stat.rutgers.edu/~madigan/BBR/

Figure 1: LLR ($\triangle$) and IC ($\bullet$) results for all pairs of rhabdomyolysis with antidepressants reported at least five times.

of an ordinary logistic regression model is an adjusted log odds ratio, i.e. a measure of association between the dependent and the explanatory variable that is adjusted for the other explanatory variables. Therefore a positive β estimate in an LLR model can be seen as a shrunk *and* adjusted log odds ratio, indicating an excessive relative reporting rate. As the IC shrinkage yields an appropriate number of new highlighted drug–ADR pairs in screening Vigibase, the LLR shrinkage was tuned to yield approximately the same number of highlighted drug–ADR pairs for LLR as for the IC.

To facilitate graphical comparison between IC and β values, we transform the IC to the natural logarithmic scale used for the β values, and use this modified IC in graphs throughout the paper: $IC^n = \log(2) * IC$.

The data on which all analyses presented in this paper are based was extracted from Vigibase at the end of the second quarter 2007, and included 3,867,085 reports from 82 countries. Retrospective evaluations over time were based on recreations of earlier versions of the data set, from the data available at the end of the second quarter 2007.

1.3 Potential Benefits of Regression

The IC and other bivariate measures [2] that rely on a cross-classification of the reports solely according to one drug and one ADR of interest are intrinsically unable to take into account the influence of all other covariates. Regression, on the other hand, estimates the effect on the reporting of an ADR of interest from a specific drug, conditional on the reporting of all other drugs. Thus, regression is superior in a pure theoretical sense.

In ADR surveillance, two specific limitations of the bivariate measures have been identified: *masking* [3] and *confounding by co-reported drugs* [10]. Masking is a phenomenon whereby the background relative reporting rate[2] of an ADR is distorted by massive reporting on the ADR with a particular drug or drug group. This could make a drug–ADR pair appear not to be excessively reported, and delay or completely hinder its highlighting for clinical review. In theory, the multivariate approach to estimating the intercepts of the regression models should make LLR less sensitive to masking. Consider for example the ADR rhabdomyolysis, a muscle tissue degenerative reaction. Following media attention,

---

[2]In the interest of readability we shall from here on write *background rate* instead of *background relative reporting rate*.
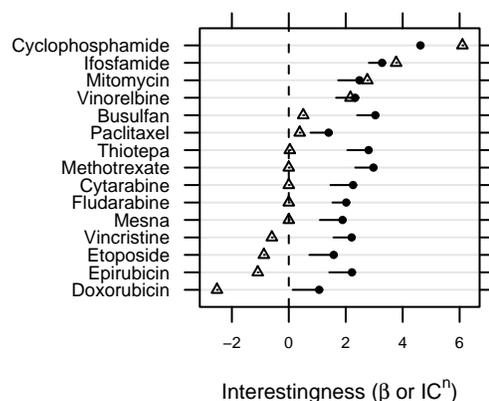

Figure 2: LLR ($\triangle$) and IC ($\bullet$) results for all pairs of haemorrhagic cystitis with anticancer drugs and mesna reported at least five times.

massive number of reports were entered into Vigibase of a particular drug, cerivastatin, with rhabdomyolysis. Rhabdomyolysis is listed on 38 % of all cerivastatin reports, to be compared to 0.3 % overall in the database. Figure 1 displays β and IC^n values for rhabdomyolysis paired with a group of drugs therapeutically very distinct from cerivastatin, antidepressants. A clear masking effect is visible, as β values are generally higher, and as 6 of the 15 drug–ADR pairs are highlighted only by LLR.

Confounding by co-reported drugs could occur when the reporting of one or more drugs correlates with the reporting of both the drug and the ADR of interest. The situation where both correlations are positive is the most well-known [10]. It could lead to a false suspicion that a drug is associated to an ADR when it is another drug, frequently co-reported with the drug of interest, which in fact is associated with the ADR. As the bivariate measures do not take into account such influence from other covariates, there are theoretical arguments why LLR should perform better than the IC in cases of confounding by co-reported drugs. Consider the example of anticancer drugs, which are often given in multi-drug regimens, so that there should be a risk of confounding by co-reporting. Figure 2 displays β and IC^n values for the ADR haemorrhagic cystitis, i.e. inflammation of the urinary bladder with bleeding, paired with anticancer drugs and mesna. There is good agreement between the methods for the strongly linked drugs, whereas a majority of the other drugs are not highlighted by LLR with this ADR. The most likely explanation is adjustment for confounding by co-reporting. An example is mesna, which is given together with the top-ranked drugs cyclophosphamide and ifosfamide to *prevent* haemorrhagic cystitis. It had 12 reports with the ADR, 10 of which also listed either cyclophosphamide or ifosfamide. Thus, the IC falsely highlights mesna with haemorrhagic cystitis whereas LLR does not.

1.4 Aim

The aim of this paper is to demonstrate how LLR can be used for pattern discovery in ADR surveillance data, and to compare our proposed method to the IC.

Table 1: Cross-classification of drug–ADR pairs according to whether they were highlighted (+) or not (−) by LLR and the *IC*, respectively.

	LLR+	LLR−			LLR+	LLR−
IC+	81,734	20,172		IC+	125	36
IC−	20,294	626,935		IC−	33	540
a. All drug–ADR pairs				**b. Nevirapine pairs**		

2. LARGE-SCALE SCREENING

In order to study systematic differences between LLR and the *IC* in their respective sets of highlighted drug–ADR pairs, both methods were used to screen the entire database. For LLR this amounts to fitting more than 2,000 models, each with almost 14,000 explanatory variables and in the order of 4 million data points.

The *IC* contrasts the relative reporting rate of an ADR given a drug of interest to a crude background rate of the ADR, namely the number of reports on the ADR relative to the total number of reports in the database. LLR uses a background rate that is a function of the estimated intercept, as implied in Section 1.3 in the context of masking. To investigate the impact of the methods' different background rates in general, and specifically how commonly LLR unmasks drug–ADR pairs relative to the *IC*, an additional screening of the database was made using a modified LLR (mLLR). For mLLR fixed intercepts were used, each corresponding to the crude background rate used by the *IC* for the ADR of interest.

Results from the screen are presented in Table 1a. A first observation is the overall high concordance between the methods. Further, 10,578 (52 %) of the 20,294 drug–ADR pairs uniquely highlighted by LLR were done so because of the difference in background rates, as they were not highlighted by mLLR. For all of these drug–ADR pairs the LLR background rate was lower than the crude. They constitute a large group that was apparently unmasked by LLR compared to the *IC*. It was also observed that whereas the *IC* requires at least three reports on a drug–ADR pair for it to be highlighted, the corresponding LLR number was four reports. This was an important difference in practice, explaining 9,060 (45 %) of the 20,172 unique *IC* highlights. The difference in background rates was, however, not particularly important among the unique *IC* highlights, as mLLR only highlighted 577 (3 %) of those. It was an unexpected finding that for 424 of those 577 drug–ADR pairs, the LLR background rate was lower than the crude.

For computational reasons, the potential importance of confounding by co-reporting in explaining differences between the methods was studied for a single drug, the anti-HIV drug nevirapine. Like the anticancer drugs, anti-HIV drugs are regularly prescribed in multi-drug regimens. For all drug–ADR pairs including nevirapine, the results of LLR and the *IC* were compared to LLR run on a modified data set where co-reported drugs were removed from nevirapine reports. This artificial method should approximate the use of LLR models that do not adjust for co-reported drugs.

Results are presented in Table 1b. The adjustment for co-reported drugs in LLR was an important explanation in the group of drug–ADR pairs uniquely highlighted by the *IC*, as 19 of 36 (53 %) were highlighted also by LLR run on the modified data set. The impact was much smaller for unique
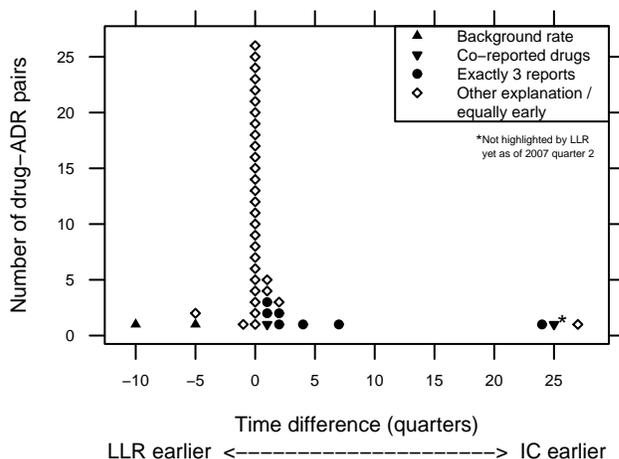

Figure 3: Difference in time required for LLR and the *IC*, respectively, to first highlight drug–ADR pairs in the retrospective analysis. Explanations are given where one method highlighted a drug–ADR pair earlier than the other method. Two drug-ADR pairs were not highlighted by either method, and are therefore not included in the graph.

LLR highlights, where 4 of 33 (12 %) of the drug–ADR pairs were not highlighted by the artificial method.

3. RETROSPECTIVE ANALYSIS

In ADR surveillance it is crucial to detect emerging issues as early as possible [1]. Therefore we compared the methods' capability of retrospectively highlighting established drug safety concerns. For the purpose of this analysis, we used a data set constructed by Lasser et al. [7] listing all withdrawals of, and major side effect warnings issued for, prescription drugs approved in the USA between 1975 and 2000. For each unique pair of a drug and a warning text or withdrawal reason, the warning text or withdrawal reason was mapped to an ADR term used in Vigibase. We determined which quarter each of these drug–ADR pairs was first highlighted (if at all). Drug–ADR pairs highlighted earlier by one method than the other were further evaluated in the quarter of the first highlight, to see whether the discrepancy could be explained by differences in number of reports, background rates or the impact of co-reported drugs, as described in Section 2.

There were 80 unique pairs of drugs and warning texts or withdrawal reasons listed in the reference. For 17 of these, the warning texts exactly matched an ADR term used in Vigibase, and for 28 of the remaining an appropriate substitute term could be found. Thus 45 drug–ADR pairs were included in the retrospective analysis.

The results are given in Figure 3. The main finding is that the vast majority of all drug–ADR pairs were highlighted by both methods. One drug–ADR pair was highlighted by the *IC* but not by LLR, and two drug–ADR pairs were not highlighted by either method. 26 of the 43 highlighted drug–ADR pairs were highlighted in the same quarter by both methods. 13 drug–ADR pairs were highlighted earlier by the *IC* than by LLR. 7 of these were highlighted by the *IC* based on just three reports, which explains why LLR did not highlight them in that same quarter. For two of the

drug–ADR pairs highlighted earlier by the *IC*, adjustment for co-reporting delayed the highlighting of the association by LLR. Four drug-ADR pairs were highlighted earlier by LLR than by the *IC*. Two of these are explained by a lower regression-based background rate compared to the crude.

4. DISCUSSION AND CONCLUSIONS

Our results show that, despite a considerable overlap, shrinkage logistic regression exemplified by LLR offers advantages of substantial practical importance compared to bivariate pattern discovery exemplified by the *IC*, in ADR surveillance. Our large-scale screenings of Vigibase using LLR and the *IC* demonstrate that unmasking is a very common explanation for why a drug–ADR pair is highlighted only by LLR; and similarly that co-reporting is important in explaining unique *IC* highlights. For the latter, the difference in the number of reports required for a highlight was also of substantial importance. These findings were coherent with those of the retrospective analysis, where LLR was found to detect some established drug safety issues earlier than the *IC*, in two instances thanks to lower background rates. However, LLR also failed to highlight some issues as early as the *IC*.

The results are dependent on the specific shrinkage and decision rule for highlighting used for LLR. A potential improvement would be to tune the shrinkage with respect to predictive accuracy, e.g. by using cross-validation. However, the pragmatic approach of calibrating LLR to highlight the same number of drug–ADR pairs as the *IC* is sensible from the perspective of the comparison itself, as the *IC* shrinkage is similarly not tuned with respect to predictive accuracy.

Given the complexity of the data with its multitude of biases and non-random missingness, transparency is a key property of any good pattern discovery method used in ADR surveillance. An example where the empirical basis for estimated LLR parameters is fairly transparent is unmasking: Lower background rates in the LLR model relative to the crude can often be attributed to one or several drugs with very large numbers of reports on the ADR of interest, as seen for rhabdomyolysis in Section 1.3. An example of the opposite situation are the drug–ADR pairs in Section 2 for which the estimated β actually *increased* in the modified LLR model with a fixed *higher* background rate. This emphasizes the complexity of interpreting parameter estimates from a regression model with more than 10,000 potential explanatory variables.

A key area of further research is to evaluate the use of shrinkage regression in other large-scale pattern discovery applications. As for ADR surveillance, possible extensions of LLR would be to include other explanatory variables, such as demographic factors (e.g. patient age and sex), or to include variables corresponding to second or higher order interactions.

In conclusion, LLR offers practical advantages over the *IC* and other bivariate methods [2] in terms of unmasking and adjustment for confounding by co-reported drugs. Further, it highlighted for clinical review some established drug safety issues earlier than the *IC*. However, since it is not always easy to identify the empirical basis for a given β estimate in a fitted LLR model, and as LLR failed to identify some established drug safety concerns as early as the *IC* in the retrospective analysis, LLR should be seen as a complement rather than as a replacement for bivariate measures.

6. REFERENCES

[1] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998.

[2] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item associations. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 67–76, New York, NY, USA, 2001. ACM.

[3] S. J. W. Evans. Statistics: analysis and presentation of safety data. In J. Talbott and P. Waller, editors, *Stephen's detection of new adverse drug reactions*, pages 301–328. Wiley & Sons, Chichester, England, 2004.

[4] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale Bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[5] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale l(1)-regularized logistic regression. *Journal of Machine Learning Research*, 8:1519–1555, 2007.

[6] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

[7] K. E. Lasser, P. D. Allen, S. J. Woolhandler, D. U. Himmelstein, S. M. Wolfe, and D. H. Bor. Timing of new black box warnings and withdrawals for prescription medications. *Journal of the American Medical Association*, 287(17):2215–2220, 2002.

[8] G. N. Norén, A. Bate, R. Orre, and I. R. Edwards. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine*, 25(21):3740–3757, 2006.

[9] G. N. Norén, R. Orre, A. Bate, and I. R. Edwards. Duplicate detection in adverse drug reaction surveillance. *Data Mining and Knowledge Discovery*, 14(3):305–328, 2007.

[10] P. Purcell and S. Barty. Statistical techniques for signal generation - The Australian experience. *Drug Safety*, 25(6):415–421, 2002.

[11] S. K. Shevade and S. S. Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[12] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.