Mathematical Statistics
Stockholm University

Statistical methods for large scale exploratory analysis of post-marketing drug safety data

G. Niklas Norén

Research Report 2005:9
Licentiate thesis

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2005:9**,
http://www.math.su.se/matstat

Statistical methods for large scale exploratory analysis of post-marketing drug safety data

G. Niklas Norén*

August 2005

Abstract

The aim of this thesis is to propose computationally feasible statistical methods for improved exploratory analysis of the complicated data sets involved in post-marketing drug safety monitoring. We implement an extended hit-miss model for duplicate detection in the WHO drug safety database and demonstrate its effectiveness on real world data. We propose improved credibility interval estimates, a Mantel-Haenszel type of adjustment for confounding variables and an extension to higher orders for the IC measure of association, which is in routine use to screen the WHO database for interesting quantitative associations. Finally, we describe how case based imprecision estimates for Bayes classifiers may be used to improve performance under asymmetrical loss functions, with a possible application in identifying series of case reports that are related to important drug safety problems.

KEY WORDS: Exploratory analysis, knowledge discovery, data mining, hit-miss model, observed-to-expected ratio, Bayes classifiers, Bayesian bootstrap

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: noren@math.su.se.

Acknowledgements

I would like to express my gratitude to all those who have provided support and encouragement during the work that has lead up to this licentiate thesis: Professor Rolf Sundberg, my main supervisor, for inspiration and advice and for sharing his expert knowledge in applied and theoretical mathematical statistics; Professor Ralph Edwards, my supervisor at the Uppsala Monitoring Centre, for providing the overall vision for our work, for helping me to broaden my views in the area of pharmacovigilance and for teaching me about the "So what?" filter in scientific publication; Andrew Bate for day to day support and advice, and for many challenging discussions on research related issues; all my other colleagues at the Uppsala Monitoring Centre, and in particular the members of the R&D team: Erik Swahn, Jonathan Edwards, Malin Ståhl, Sven Purbe and Johan Hopstadius, as well as Roland Orre at NeuroLogic; all PhD students and all faculty at Mathematical Statistics, Stockholm University, for welcoming me to the group and for providing a stimulating research environment; finally my parents Hans-Göran and Christina Norén for their everlasting support and encouragement and my girlfriend Minna for making it all worthwhile.

Uppsala, August 2005,
Niklas Norén

List of papers

I Norén GN, Orre R, Bate A. A hit-miss model for duplicate detection in the WHO drug safety database. To appear in the Proceedings of the Eleventh ACM SIGKDD International Conference on **Knowledge Discovery and Data Mining**, Chicago, Illinois, USA, August 21-24, 2005.

II Norén GN, Bate A, Orre R. Extending the methods used to screen the WHO drug safety database towards the analysis of complex associations and improved accuracy for rare events. **Submitted for publication**.

III Norén GN, Orre R. Case based imprecision estimates for Bayes classifiers with the bayesian bootstrap. **Machine Learning** 58, 79-94, 2005.

1 Introduction

The analysis of spontaneous adverse drug reaction (ADR) report databases is one of the most important tools for early discovery of drug safety problems not identified prior to marketing (Rawlins 1988). Because ADR case reports are based on actual clinical practice, the chance is greater than in pre-marketing studies to discover ADRs that are due to drug interactions, occur only after extended periods of use or affect groups that are typically excluded from clinical trials (such as children or pregnant women). The large cohort exposed also makes rare ADRs easier to pick up.

Spontaneous reporting data sets are often large and complex – the WHO ADR database described in Section 2 contains more than 3 million individual case reports. While manual evaluation of this data set may certainly provide useful insight into specific issues, broader studies of more open ended questions require efficient computational methods, as discussed in Section 3. These methods are more than just a computational necessity, however. A method referred to as IC analysis (IC is short for Information Component) has been developed to identify interesting quantitative associations between rare events in large data sets (Bate et al. 1998) and has been in routine use on the WHO drug safety database since 1998. Hypotheses related to potential drug safety problems first highlighted with this method are routinely communicated to the drug safety community, and some have been published even in the mainstream medical literature (Coulter et al. 2001, Sanz et al. 2005).

The aim of this thesis is to propose improvements and extensions to the computational statistical methods for exploratory analysis of post-marketing drug safety data. Paper **I** proposes a new method for automated duplicate detection based on the hit-miss model for statistical record linkage (matching records across data sets) introduced by Copas and Hilton (1990), with improvements to handle numerical record fields and a method to compensate for correlations between record fields. The method is implemented for the WHO database and demonstrated to be useful in real world duplicate detection. Paper **II** proposes improved credibility interval estimates, a Mantel-Haenszel type of adjustment for confounding variables and an extension to higher orders for the IC measure of association used to screen the WHO database for interesting quantitative associations. Paper **III** introduces a Bayesian bootstrap method for estimating the uncertainty in Bayes classification and improving performance under asymmetrical loss, with a possible application to post-marketing drug safety data as discussed in Section 4.3.

2 The WHO drug safety database

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden holds the world's largest collection of reports on suspected ADR incidents from drug substances after they have been introduced on the market. This is an observational data set that consists of spontaneous reports provided by health professionals around the world upon the observation of suspected ADR incidents in clinical practice. Reports are routinely forwarded to Uppsala from national drug safety centres in the 76 member countries of the WHO Programme for International Drug Monitoring, and the oldest reports in the data set date back to 1968.

2.1 Size

The most striking feature of the WHO database is perhaps its size. The database currently contains over 3.5 million case reports, with an additional approximately 200,000 being added each year. The number of variables of potential interest is massive: drug substances and ADR terms alone make for over 17,000 binary variables (roughly 15,000 drug substances and 2,000 ADR terms). In addition, studies may involve variables such as patient age and gender, reporting country, prescription date, onset date, dosage and outcome. In order to handle the large number of case reports and variables involved, any method for exploratory analysis of this data set must be both computationally robust and efficient. One must also account for the sparsity of data: with over 3 million reports in total, it may seem a paradox that lack of data should ever be a problem, but the large number of reports is balanced by an equally large number of possible variables. There are, for example, in the order of 30 million possible pairs of one drug substance and one ADR term. Out of these, around 600,000 occur together on case reports in the database, with a median joint count of 2 and a mere 15% being co-reported more than 10 times (Norén 2002).

2.2 Characteristics

The case reports in the WHO database refer to *suspected* ADR incidents, and some reported events will in reality have been coincidental or perhaps due to concomitant medication or the underlying disease. In addition, far from all ADR incidents that occur are reported, and variations in the degree of under-reporting between different events also make it difficult to interpret raw numbers of reports. Most modern methods for quantitative analysis of post-marketing drug safety data compare the reporting of specific pairs of events to a reference based on marginal relative frequencies of the events in the database as a whole (Bate et al. 1998, DuMouchel 1999, Evans et al.

2001, Egberts et al. 2002). A benefit of this is that variations in the marginal reporting rates (due to for example regulatory requirements, the severity of the reaction or the time on the market for the drug) may be automatically compensated for. Reporting rate variations that affect specific combinations of events are more difficult to account for, however. If for example, attention in the media or in the scientific community leads to increased reporting of a specific drug-ADR pair, this will be very difficult to account for in the analysis. Another example of problematic relative over-reporting is when an individual health professional is responsible for several reports in a specific case series, as in one of the examples in the empirical studies in **I**. The use of the database as the reference in estimating strength of association is based on the assumption that it well reflects the general reporting rates for different events and it is important to consider potential violations of this assumption in the interpretation of estimated strengths of association: unusually large numbers of reports on specific case series may, for example, lead to misleadingly large marginal frequencies for the terms involved. This would reduce their estimated strength of association with other events, and may lead to true problems being missed.

There are several sources of heterogeneity in the WHO database, which are important to be aware of in quantitative investigations. The range of available drug substances, populations at risk, reporting behaviour and regulations may vary both between countries and over time. In addition, not all reports in the data set have been spontaneously submitted but a small proportion come from intensive monitoring programs where relative reporting rates are higher. It may also be argued that due to fundamental differences in how vaccines are administered compared to other drug substances, reports on adverse reactions to vaccines should be analysed separately. The problem in practice is that the information necessary to identify different report types is often not available for all case reports.

2.3 Data quality

The quality of individual case reports is highly variable. Whereas some reports are highly detailed and accurate, others are incomplete, inconsistent or incorrect. This is of great importance in the clinical review of case series, but little work has been done to automatically account for it in the quantitative analysis. Missing data is an important problem primarily in specialised studies involving record fields other than drug substances and ADR terms (which together with the reporting country are rarely completely missing from a case report). Strategies for handling missing data thus become important primarily in screening for risk factors or in carrying out subgroup analyses. The problem with duplicate reports considered in **I** differs from other data quality problems in that it does not relate to the quality of a single report, but to the quality of the data set as a whole. Even upon the identification of a pair of duplicate reports it is not obvious how to proceed. Should one report be flagged or perhaps removed from the data set (and if so, which one)?

3 Exploratory analysis

The aim of exploratory studies is not to draw final conclusions but to generate and refine hypotheses with respect to the content and the nature of a data set. Quite naturally, exploratory investigations tend to be more open ended and have less rigidly defined objectives than confirmatory studies, but they do require at least the *types* of hypotheses to be specified beforehand (Hand 1998). The identification of an appropriate statistical method for a given study objective is critical to successful exploratory data analysis, and requires both insight into the nature of data and proficiency in applied statistical inference. Methods for exploratory analysis of large databases may be referred to as exploratory data analysis (Hand et al. 2001, Tukey 1977), knowledge discovery (Fayyad et al. 1996) or data mining (Hand 1998) depending on the context. For a discussion on how they relate, see for example Elder and Pregibon (1996), Glymour et al. (1997), Hand (1999), Breiman (2001) and Hastie et al. (2001).

Exploratory studies often lead to results that relate not to the primary study objective, but to underlying properties of the data or of the data collection process. In practice, data cleaning and analysis are often intertwined, so that data analysis results lead to improved data quality, which in turn allows for more accurate data analysis. In addition, data analysis itself is often carried out in an iterative fashion of stepwise refinement. It is rarely possible to specify at the outset of a large exploratory study, a detailed data analysis method appropriate for all possible questions and patterns under study.

As an illustration of the need for stepwise refinement strategies in exploratory data analysis, consider the adjustment for confounding variables in large scale screening for quantitative associations in post-marketing drug safety data. The range of possible confounders in the WHO database includes patient age, patient gender, reporting country, reporting date and concomitant medication. It has been suggested that exploratory studies of post-marketing drug safety data should routinely adjust for possible confounding variables (Lillienfeld et al. 2003), but simultaneous adjustment for the possible presence of all 14,000 drug substances would lead to an extreme $2^{14,000}$ different strata. Even if concomitant medication is ignored, simultaneous adjustment for gender (3 groups), age (10 groups), country (75 groups) and reporting quarter (144 groups) would yield 324,000 different strata (out of which over 30,000 contain at least one case report in the current version of the WHO database). Such large numbers of strata are not only a computational challenge, but problematic also from a methodological point of view. A sensible stepwise refinement approach may be to do the initial screen based on unadjusted (or carefully adjusted) estimates, and follow up highlighted associations by automated confounder detection and adjustment. The main challenge is how to handle downward confounding (when true problems are missed due to under-estimated strength of association), and more research into the tendency of different covariates to confound drug-ADR associations

in the WHO database is necessary before an optimal strategy for automated confounder adjustment can possibly be identified.

3.1 Application to post-marketing drug safety data

In screening massive post-marketing drug safety data sets for interesting associations involving a large number of possible events, the great number of hypotheses considered makes the exploratory approach a necessity. For spontaneous reporting data, the inherent data biases and variable quality of case reports further motivates an exploratory attitude. The emphasis on hypothesis generation and refinement applies throughout this thesis: the aim of the algorithm in **I** is to highlight *likely* duplicates for manual review and the aim of the IC analysis methodology in **II** (and the Bayes classifiers in **III** if implemented for the WHO database as proposed in Section 4.3) is to highlight potential drug safety issues for clinical evaluation. The focus on hypothesis generation does not reduce the need for effective statistical algorithms: the amount of resources available for clinical review is limited and must be managed wisely – every false lead followed up may be at the expense of an undiscovered true problem. At the same time, false negatives are certainly more problematic than false positives in that they never reach the clinical review.

There is a wide range of possible applications for computational statistical methods in exploratory analysis of post-marketing drug safety data sets, including:

- Duplicate detection and record matching (see **I**)
- Screening for pairwise and higher order quantitative associations between drug substances, ADR terms and other types of events (see **II**)
- Classification of case series with respect to the probability that they will be considered interesting in the clinical review (see Section 4.3)
- Clustering of case reports based on the reported ADR terms, with the aim of identifying previously unknown syndromes of ADRs (see Orre et al. (2005))
- Clustering of drug substances with respect to their ADR profiles, in order to identify groups with similar pharmacodynamical properties
- Automated quality grading of case reports and methods to account for this in quantitative studies
- Methods for automated identification of potential confounders and risk factors

4 Comments to the papers

The papers included in this thesis relate to different aspects of the knowledge discovery process. The overall aim is to extract useful information from data that was not collected primarily for the purpose of data analysis. A technical link between the papers is their use of ratios on the following form:

$$\frac{P(x,y)}{P(x)P(y)} = \frac{P(y \mid x)}{P(y)} \tag{4.1}$$

Such ratios relate the joint probability for two events x and y to the corresponding probability under the assumption that the two events are independent. Based on data, a natural estimate for (4.1) is the oberved-to-expected ratio of the relative frequencies:

$$\frac{O_{xy}}{E_{xy}} = \frac{f(x,y)}{f(x)f(y)} \tag{4.2}$$

The reference to E_{xy} as the "expected" may be confusing since it is not an expected value in the statistical sense but a natural estimate for the joint probability under the independence model. Moreover the distinction between the theoretical quantity and its estimate is often not made clear, and in this thesis I loosely refer to both (4.1) and (4.2) as observed-to-expected ratios.

In **I**, these ratios are used to compensate for correlations between record fields in hit-miss model duplicate detection, in **II** they are used as measures of association between events and in **III** they correspond to the contribution from different explanatory variable to the Bayes classifier output probabilities. They also serve as weights in a Hopfield neural network that has been used to generate hypotheses of complex associations between groups of ADR terms (syndromes) in post-marketing drug safety data sets (Orre et al. 2005).

4.1 Paper I

The immediate aim of the duplicate detection method in **I** is to improve data quality, which in the end should allow for more accurate data analysis (although under certain circumstances, record matching may be considered data analysis in its own right). Clearly, the identification of duplicate case reports in the WHO drug safety database can be expected not only to make quantitative investigations more reliable but also to facilitate clinical review.

The duplicate detection method in **I** is based on the hit-miss model for statistical record linkage (matching records across data sets) introduced by Copas and Hilton (1990). Based on a probabilistic model for how discrepancies between related database records occur, this duplicate detection method essentially attributes a weight for each record field depending on whether the

two records of interest match, mismatch or lack information in this field. Matches on rare events receive higher weights than matches on more common events, and the penalty for mismatching information depends on how common mismatches are for that particular record field in the training data. The total match score for a pair of database records is found by adding together the weights for the different record fields (with an adjustment for potential correlations).

In the empirical study reported on in Section 3.1 of **I**, three case reports other than the known duplicate received unexpectedly high match scores together with one particular case report in the test data set (see Table 6 of **I**). Manual inspection of the full case records in the WHO database did not allow these to be dismissed as false positives, and follow up information has been requested from the FDA in order to determine whether they are in fact unidentified true duplicates. As discussed in Section 4 in **I**, the three confirmed false positives in the experiment on Norwegian data relate to a group of case reports provided by one dentist on the same drug-ADR pair. Unexpectedly (given that they relate to different patients), all three case reports have the same listed onset date. This may indicate a data quality problem and illustrates how, in the knowledge discovery process, data analysis may allow for improved data cleaning as discussed in Section 3.

4.2 Paper II

The aim of the extensions and improvements to the IC analysis methodology proposed in **II** is to allow for more sophisticated and reliable screening for quantitative associations in the WHO database. Some comments with respect to the use of the IC measure of association may be appropriate. The IC is defined as the logarithm of (4.1):

$$IC_{xy} = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{P(y \mid x)}{P(y)} \tag{4.3}$$

The transformation to the logarithmic scale allows for an interpretation of the IC as a residual under the independence model: $\log_2 O_{xy} - \log_2 E_{xy}$ and IC analysis as a form of outlier detection. There are efficient shrinkage estimators for the IC which are useful in exploratory analysis of large and complicated data sets and these are studied in detail in **II**. Their main advantage is that they are less sensitive to low values in the denominator than raw observed-to-expected ratios, and this reduces the vulnerability to spurious associations. Figure 1 illustrates how the difference between IC shrinkage estimates and raw log-observed-to-expected ratios changes with increasing E_{xy}.

Strength of association estimates based on the observed-to-expected ratio are most useful for rare events, such as the presence of specific drug substances and ADR terms on case reports in the WHO database. Studies that involve

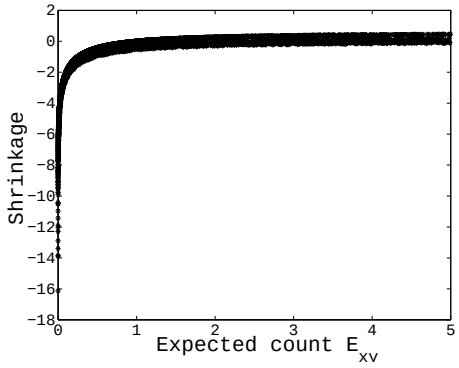


Figure 1: Difference between the IC shrinkage estimate and the raw log-observed-to-expected ratio for 10,000 randomly selected drug-ADR pairs with at least one case report, sorted by E_{xy}.

other types of events or specific data subsets may motivate the use of for example the log-odds ratio instead. The problem with the observed-to-expected ratio is that it contrasts $P(y \mid x)$ to $P(y) = P(y \mid x)P(x) + P(y \mid x^c)P(x^c)$, which as $P(x)$ approaches one will be determined largely by the former. The ratio of $P(y \mid x)$ to $P(y)$ will consequently be close to 1 for $P(x)$ near 1, regardless of how much $P(y \mid x)$ deviates from $P(y \mid x^c)$.

4.3 Paper III

The Bayesian bootstrap method for Bayes classifiers in **III** was evaluated on standard data sets from the UCI Machine Learning Repository, but could potentially be used to identify those case series in the WHO database that are most likely to relate to real drug safety problems. Useful explanatory variables for such an implementation may include case series characteristics such as the total number of reports, their quality and geographic spread. Under the assumption that false negatives (missed true problems) are more problematic than false positives (extra work for the clinical experts), the loss functions involved will be asymmetrical, and the Bayesian bootstrap method should improve performance.

Bayes classifiers can be re-expressed in terms of ratios of the form in (4.1). The naive Bayes classifier is based on the assumption that the explanatory variables are independent conditional on the class: $P(x_1, \ldots, x_m \mid y) = P(x_1 \mid y) \cdot \ldots \cdot P(x_m \mid y)$. For a given set of explanatory variable values

$x_1, \ldots, x_m$, it can be re-expressed as:

$$\begin{aligned} P(y_j \mid x_1, \ldots, x_m) &= \frac{P(x_1, \ldots, x_m \mid y_j) \cdot P(y_j)}{P(x_1, \ldots, x_m)} \\ &\approx \frac{P(x_1 \mid y_j) \cdot \ldots \cdot P(x_m \mid y_j) \cdot P(y_j)}{P(x_1, \ldots, x_m)} \\ &\propto \frac{P(x_1 \mid y_j)}{P(x_1)} \cdot \ldots \cdot \frac{P(x_m \mid y_j)}{P(x_m)} \cdot P(y_j) \\ &= \frac{P(x_1, y_j)}{P(x_1)P(y_j)} \cdot \ldots \cdot \frac{P(x_m, y_j)}{P(x_m)P(y_j)} \cdot P(y_j) \end{aligned} \tag{4.4}$$

or on the logarithmic scale:

$$\log_2 P(y_j \mid x_1, \ldots, x_m) = \log_2 \frac{P(x_1, y_j)}{P(x_1)P(y_j)} + \ldots + \log_2 \frac{P(x_m, y_j)}{P(x_m)P(y_j)} + \log_2 P(y_j) \tag{4.5}$$

IC shrinkage estimates could be used instead of raw observed-to-expected ratios in (4.5) to reduce the sensitivity to spurious associations between explanatory and response variables in training data. In addition, the re-expression allows for an interpretation of the IC measure of association in terms of the predictive strength of one event on another. The replacement of $P(x_1, \ldots, x_m)$ by $P(x_1) \cdot \ldots \cdot P(x_m)$ above was not used in **III** because the naive Bayes assumption of independence between explanatory variables only holds conditional on class membership. In practice, these quantities are independent of class membership and cancel in the normalisation of the estimated class probabilities to sum to 1.

In **III**, the marginal class probabilities $P(y_j)$ were estimated based on the proportion of instances from each class in the available training data. This is appropriate when training data is a representative sample from the population to which the classifier is to be applied in the future. If, on the other hand, the composition of training data does not necessarily represent future observations, then $P(y_j)$ must be based either on external data relevant to the population of interest or on prior knowledge. The adjusted Bayesian bootstrap approach can be modified to accommodate this, by replacing the numbers of cases for each class in training data $\{n_{y_1}, n_{y_2}, \ldots\}$ (see Table 1 in **III**) by the corresponding numbers in the external data set (or pseudo-counts if the class probabilities are based on prior knowledge).

5 Summary

Computationally feasible statistical methods may facilitate the exploratory analysis of post-marketing drug safety data, allowing for large scale hypothesis generation and refinement. In this thesis, new methods are introduced to improve data quality and allow for more sophisticated data analysis:

- The hit-miss model is extended and adapted for duplicate detection in the WHO drug safety database, with good performance demonstrated on real world data
- The IC analysis methodology for screening post-marketing drug safety data for quantitative associations is extended to allow for more accurate credibility interval estimates, to a Mantel-Haenszel type of adjustment for confounding variables and to the analysis of higher order quantitative associations
- A Bayesian bootstrap approach is proposed for the estimation of uncertainty in Bayes classifier predictions and for improvement of performance under asymmetrical loss

References

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

Breiman, L.: 2001, Statistical modeling: the two cultures, *Statistical science* **16**(3), 199–231.

Copas, J. and Hilton, F.: 1990, Record linkage: statistical models for matching computer records, *Journal of the Royal Statistical Society: Series A* **153**(3), 287–320.

Coulter, D. M., Bate, A., Meyboom, R. H., Lindquist, M. and Edwards, I. R.: 2001, Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study, *British Medical Journal* **322**(7296), 1207–1209.

DuMouchel, W.: 1999, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting systems, *American Statistician* **53**, 177–202.

Egberts, A. C., Meyboom, R. H. and van Puijenbroek, E. P.: 2002, Use of measures of disproportionality in pharmacovigilance: three Dutch examples, *Drug Safety* **25**(6), 453–458.

Elder, J. F. and Pregibon, D.: 1996, A statistical perspective on knowledge discovery in databases, *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 83–113.

Evans, S., Waller, P. and Davis, S.: 2001, Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports, *Pharmacoepidemiology and Drug Safety* **10**(6), 483–486.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: 1996, From data mining to knowledge discovery: an overview, *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 1–34.

Glymour, C., Madigan, D., Pregibon, D. and Smyth, P.: 1997, Statistical themes and lessons for data mining, *Data Min. Knowl. Discov.* **1**(1), 11–28.

Hand, D. J.: 1998, Data mining: Statistics and more?, *The American Statistician* **52**, 112–118.

Hand, D. J.: 1999, Statistics and data mining: intersecting disciplines, *SIGKDD Explor. Newsl.* **1**(1), 16–19.

Hand, D. J., Mannila, H. and Smyth, P.: 2001, *Principles of Data Mining*, MIT Press.

Hastie, T., Tibshirani, R. and Friedman, J.: 2001, *The elements of statistical learning: data mining, inference and prediction*, Springer.

Lillienfeld, D., Nicholas, S., Macneil, D., Kurjatkin, O. and Gelardin, T.: 2003, Violation of homogeneity: a methodological issue in the use of data mining tools, *Drug Safety* **26**, 363–364.

Norén, N.: 2002, *A Monte Carlo method for Bayesian dependency derivation*, Master's thesis, Chalmers University of Technology.

Orre, R., Bate, A., Norén, G. N., Swahn, E., Arnborg, S. and Edwards, I. R.: 2005, A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets, *International Journal of Neural Systems* **15**(3), 207–222.

Rawlins, M. D.: 1988, Spontaneous reporting of adverse drug reactions. II: Uses, *British Journal of Clinical Pharmacology* **1**(26), 7–11.

Sanz, E. J., De-las-Cuevas, C., Kiuru, A., Bate, A. and Edwards, I. R.: 2005, Selective serotonin reuptake inhibitors in pregnant women and neonatal withdrawal syndrome: a database analysis, *The Lancet* **365**, 482–487.

Tukey, J. W.: 1977, *Exploratory data analysis*, Addison-Wesley Pub. Co.

I

A Hit-Miss Model for Duplicate Detection in the WHO Drug Safety Database

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

Roland Orre
NeuroLogic Sweden AB
Stockholm, Sweden
roland.orre@neurologic.se

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

ABSTRACT

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden, maintains and analyses the world's largest database of reports on suspected adverse drug reaction incidents that occur after drugs are introduced on the market. As in other post-marketing drug safety data sets, the presence of duplicate records is an important data quality problem and the detection of duplicates in the WHO drug safety database remains a formidable challenge, especially since the reports are anonymised before submitted to the database. However, to our knowledge no work has been published on methods for duplicate detection in post-marketing drug safety data. In this paper, we propose a method for probabilistic duplicate detection based on the hit-miss model for statistical record linkage described by Copas & Hilton. We present two new generalisations of the standard hit-miss model: a hit-miss mixture model for errors in numerical record fields and a new method to handle correlated record fields. We demonstrate the effectiveness of the hit-miss model for duplicate detection in the WHO drug safety database both at identifying the most likely duplicate for a given record (94.7% accuracy) and at discriminating duplicates from random matches (63% recall with 71% precision). The proposed method allows for more efficient data cleaning in post-marketing drug safety data sets, and perhaps other applications throughout the KDD community.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical computing; H.2.m [**Database Management**]: Miscellaneous; J.3 [**Life and medical sciences**]: Health

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

General Terms

Algorithms

Keywords

Duplicate detection, hit-miss model, mixture models

1. INTRODUCTION

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden (also known as *the* Uppsala Monitoring Centre) holds the world's largest database of spontaneous reports on suspected adverse drug reaction (ADR) incidents. Spontaneous reports are provided to pharmaceutical companies and regulatory bodies by health professionals upon the observation of suspected ADR incidents in clinical practice. The 75 member countries of the WHO Programme for International Drug Monitoring routinely forward ADR case reports submitted to their medical products agencies to *the* Uppsala Monitoring Centre. The first case reports in the WHO drug safety database date back to 1967 and as of January 2005 there are over 3 million reports in total in the data set; currently around 200,000 new reports are added to the database each year.

While the analysis of spontaneous reporting data is one of the most important methods for discovering previously unknown safety problems after drugs are introduced on the market [16], it is sometimes impaired by poor data quality [11], and in particular the presence of duplicate case reports. Quantitative methods are important in screening spontaneous reporting data for new drug safety problems [1], and may highlight potential problems based on as few as 3 case reports on a particular event, so the presence of just 1 or 2 duplicates may severely affect their efficacy. While there is a general consensus that the presence of duplicates is a major problem in spontaneous reporting data, there is a lack of published research with respect to the extent of the problem. A study on vaccine ADR data quoted proportions of around 5% confirmed duplicates [14]. However, at times the frequency may be much higher: in a recent review of suspected quinine induced thrombocytopenia, FDA researchers identified 28 of the 141 US case reports (20%) as duplicates [6].

There are at least two common causes for duplication in post-marketing drug safety data: different sources (health professionals, national authorities, different companies) may provide separate case reports related to the same event and there may be mistakes in linking follow-up case reports to earlier records. (Follow-up reports are submitted for example when the outcome of an event is discovered.) The risk of duplication is likely to have increased in recent years due to the advent of information technology that allows case reports to be sent back and forth more easily between different organisations [8], and the transfer of case reports from national centres to the WHO might introduce extra sources of error, including the risk that more than one national centre provide case reports related to the same event.

Duplicate records are typically much more similar than random pairs of records. There are however important exceptions. For example, separate case reports are sometimes provided for the same patient recorded at the same doctor's appointment when the patient has suffered from unrelated ADRs. Such record pairs may match perfectly on date, age, gender, country and drug substances, but should not be considered as duplicates. The opposite problem is illustrated by so called mother-child reports that relate to ADR incidents in small children from medication taken by the mother during pregnancy. Such record pairs differ greatly depending on whether the patient information relates to the mother or the child.

The need for algorithms to systematically screen for duplicate records in drug safety data sets is clear [5]. There are no published papers in this area, but general duplicate detection methods are available [3, 10, 12, 17]. In addition, the fundamentally similar problem of record linkage (matching records across data sets) has been studied since the 1960s [9, 13]. We have chosen to develop a duplicate detection method based on the hit-miss model for statistical record linkage described by Copas & Hilton [7]. The hit-miss model has several important beneficial properties. It imposes no strict criteria that a pair of records must fulfil in order to be highlighted as suspected duplicates, which is useful for spontaneous reporting data where errors occur in all record fields. Rather than just classifying record pairs as likely duplicates or not, the hit-miss model provides a prioritisation (scoring) with respect to the chance that a given pair of records are duplicates. This allows the number of record pairs highlighted to be adjusted depending on the resources available for manual review. While the hit-miss model punishes discrepancies it rewards matching information, which ensures that identical record pairs with very little data listed are unlikely to be highlighted for follow-up at the expense of more detailed record pairs with slight differences. Furthermore, the reward for matching information varies depending on how common the matching event is, so that for example a match on a rare adverse event is considered stronger evidence than a match on gender. The fact that most of the hit-miss model parameters are determined by the properties of the entire data set reduces the risk of over-fitting the algorithms to training data, which is very important for the WHO database, where the amount of labelled training data is limited.

The aim of this paper is to propose two new improvements to the standard hit-miss model (a model for errors in numerical record fields and a computationally efficient approach to handling correlated record fields) and to show that the adapted hit-miss model is very useful in real world duplicate detection. We fit the hit-miss model to the WHO drug safety database, and evaluate its performance on a test set of real world database records that includes a certain proportion of known duplicates.

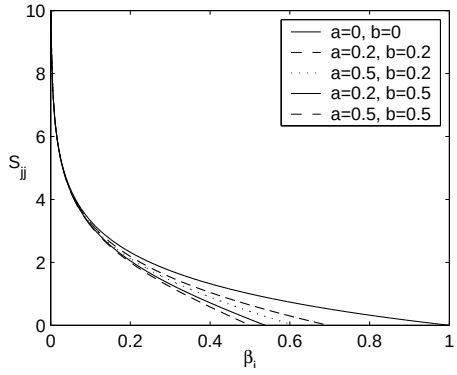


Figure 1: $W_{jj}(\beta_j)$ based on (8), for several values of a and b

2. METHODS

2.1 The hit-miss model

2.1.1 The standard hit-miss model

The hit-miss model is a probability model for how discrepancies between database records that relate to the same underlying event occur. Let $X = j$ and $Y = k$ denote the observed values on two database records for a certain record field and let p_j and p_k denote the corresponding probabilities. The joint probability for this pair of values under the independence assumption is $p_j \cdot p_k$. The hit-miss model provides an estimate p_{jk} for the same probability under the assumption that the two records are duplicates. The contribution from each record field (its weight) to the total match score is equal to the log-likelihood ratio for the two hypotheses (high values correspond to likely duplicates):

$$W_{jk} = \log_2 \frac{p_{jk}}{p_j p_k} \tag{1}$$

and the total match score is found by adding together the weights for all different record fields.

Under the hit-miss model, each observed record field value X is based on a true but unobserved event T. Observed values are assumed to be either misses, blanks or hits. Misses occur with probability a, blanks with probability b and hits with probability $1-a-b$. For a miss X is a random variable following the overall incidence of T, for a blank the value of X is missing and for a hit $X = T$.

Let $P(T = i) = \beta_i$ and let $P(X = j \mid T = i) = \alpha_{ji}$. The following holds generally under the assumption that X and Y are independent conditional on T:

$$p_{jk} = \sum_i \alpha_{ji}\alpha_{ki}\beta_i \tag{2}$$

Outcomes	Probability	Distribution
H,H	$(1-a_1-a_2-b)^2$	$\delta(d)$
H,D	$2a_1(1-a_1-a_2-b)$	$\phi(d;0,\sigma_1^2)$
D,D	a_1^2	$\phi(d;0,2\sigma_1^2)$
H,M	$2a_2(1-a_1-a_2-b)$	$f(d)$
M,M	a_2^2	$f(d)$
D,M	$2a_1a_2$	approx $f(d)$

Table 1: Outcomes of interest (H=hit, D=deviation, M=miss) in the hit-miss mixture model, together with associated probabilities and distributions for d.

Under the hit-miss model:

$$\alpha_{ji} = \begin{cases} a\beta_j & j \neq i \\ 1-b-a(1-\beta_j) & j = i \\ b & j \text{ blank} \end{cases} \quad (3)$$

and it can be shown that if $c = a(2-a-2b)$:

$$p_{jk} = \begin{cases} c\beta_j\beta_k & j \neq k \\ \beta_j\{(1-b)^2 - c(1-\beta_j)\} & j = k \\ b(1-b)\beta_k & j \text{ blank} \\ b^2 & j,k \text{ blank} \end{cases} \quad (4)$$

Based on (4):

$$P(X=j) = (1-b)\cdot\beta_j \quad (5)$$

$$P(X \text{ blank}) = b \quad (6)$$

$$P(\text{discordant pair}) = c\cdot(1-\sum_i \beta_i^2) \quad (7)$$

Thus, for a given record field, we estimate b by its relative frequency of blanks in the entire database and β_i by its relative frequency of value i among non-blanks in the entire database. c is estimated by the relative frequency of discordant pairs for this record field among non-blanks in the set of identified duplicate pairs, divided by $1-\sum_i \beta_i^2$.

(3), (4) and (5) give:

$$W_{jk} = \begin{cases} \log_2 c - 2\log_2(1-b) & j \neq k \\ \log_2\{1-c(1-\beta_j)(1-b)^{-2}\} - \log_2\beta_j & j = k \\ 0 & j \text{ or } k \text{ blank} \end{cases} \quad (8)$$

Thus, all mismatches for a given record field receive the same weight and blanks receive weight 0. It can be shown that matches on rare events receive greater weights than matches on more common events (W_{jj} decreases when β_j increases) as would intuitively be expected. The detailed behaviour of W_{jj} as a function of β_j is illustrated in Figure 1 for different values of a and b.

2.1.2 A hit-miss mixture model for errors in numerical record fields

For numerical record fields such as date and age, many types of error are more likely to yield small differences between true and observed values. If, for example, two different sources send separate case reports related to the same incident, the dates may perhaps disagree, but it is more likely that they should differ by a few days than by several years. Similarly, the registered age for patient sometimes differs from the true value, but then a small difference is more likely than a large one. At the same time, there may be other types of errors (*e.g.* typing errors) where a large numerical difference is as likely as a small one. In order to handle both possibilities, we propose a hit-miss mixture model which includes a new type of miss for which small deviations from the true value are more likely than large ones. To distinguish between the two types of misses in this model, we refer to the first type as 'misses' and the second type as 'deviations'. If T is the true, but unobserved value, then X is a random variable assumed to have been generated through a process that results in a deviation with probability a_1, a miss with probability a_2, a blank with probability b and a hit with probability $1-a_1-a_2-b$. For a deviation, X follows a $N(T,\sigma_1^2)$ distribution and for a miss, X is a random variable independent of T but with the same distribution. For a blank, the value of X is missing and for a hit, $X = T$.

For two observed numerical values $X = i$ and $Y = j$, we focus on the difference $d = j - i$. For duplicates we must distinguish between 6 possible outcomes for the hit-miss mixture model as listed in Table 1 where $\phi(d;\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 and $\delta(d)$ denotes Dirac's delta function, which has all its probability mass centred at 0. $f(d)$ denotes the probability density function for the difference between two independent random events that follow the same distribution as T, such as for example hits and misses. Under the assumption that $\text{var}(T) \gg \sigma_1^2$, the difference between a miss and a deviation also follows this distribution.

Thus, the hit-miss mixture model for the difference d between the numerical values for two duplicates can be reduced

1. Make initial guesses for the parameters $\hat{a}_1$, $\hat{a}_2$ and $\hat{\sigma}_1^2$

2. *Expectation step:* Calculate $\hat{\alpha}_1,\ldots,\hat{\alpha}_4$:
   $\hat{\alpha}_1 = (1-\hat{a}_1-\hat{a}_2-\hat{b})^2$
   $\hat{\alpha}_2 = \hat{a}_2(2-2\hat{b}-\hat{a}_2)$
   $\hat{\alpha}_3 = 2\hat{a}_1(1-\hat{a}_1-\hat{a}_2-\hat{b})$
   $\hat{\alpha}_4 = \hat{a}_1^2$
   For each observed d_i in training data, compute the probability that it belongs to each mixture component
   $\hat{\gamma}_1(d_i) = \frac{\hat{\alpha}_1\delta(d_i)}{\hat{\alpha}_1\delta(d_i)+\hat{\alpha}_2 f(d_i)+\hat{\alpha}_3\phi(d_i;0,\hat{\sigma}_1^2)+\hat{\alpha}_4\phi(d_i;0,2\hat{\sigma}_1^2)}$
   $\hat{\gamma}_2(d_i) = \frac{\hat{\alpha}_2 f(d_i)}{\hat{\alpha}_1\delta(d_i)+\hat{\alpha}_2 f(d_i)+\hat{\alpha}_3\phi(d_i;0,\hat{\sigma}_1^2)+\hat{\alpha}_4\phi(d_i;0,2\hat{\sigma}_1^2)}$
   $\hat{\gamma}_3(d_i) = \frac{\hat{\alpha}_3\phi(d_i;0,\hat{\sigma}_1^2)}{\hat{\alpha}_1\delta(d_i)+\hat{\alpha}_2 f(d_i)+\hat{\alpha}_3\phi(d_i;0,\hat{\sigma}_1^2)+\hat{\alpha}_4\phi(d_i;0,2\hat{\sigma}_1^2)}$
   $\hat{\gamma}_4(d_i) = \frac{\hat{\alpha}_4\phi(d_i;0,2\hat{\sigma}_1^2)}{\hat{\alpha}_1\delta(d_i)+\hat{\alpha}_2 f(d_i)+\hat{\alpha}_3\phi(d_i;0,\hat{\sigma}_1^2)+\hat{\alpha}_4\phi(d_i;0,2\hat{\sigma}_1^2)}$

3. *Maximisation step:* Calculate the weighted variance $\hat{\sigma}_1^2$:
   $\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n \hat{\gamma}_3(d_i)\cdot d_i^2 + \hat{\gamma}_4(d_i)\cdot d_i^2/2}{\sum_{i=1}^n \hat{\gamma}_3(d_i)+\hat{\gamma}_4(d_i)}$
   Update $\hat{a}_1$ and $\hat{a}_2$ by numerical maximisation of the total likelihood for the observed data over eligible value pairs (such that $\hat{a}_1+\hat{a}_2+\hat{b} < 1$).

4. Iterate 2-3 until convergence

Table 2: EM algorithm for the hit-miss mixture model.

to four components:

$$p_d(d) = (1-a_1-a_2-b)^2 \cdot \delta(d) + a_2(2-a_2-2b) \cdot f(d) + \\ + 2a_1(1-a_1-a_2-b) \cdot \phi(d;0,\sigma_1^2) + a_1^2 \cdot \phi(d;0,2\sigma_1^2) \quad (9)$$

For unrelated records, d follows the more simple distribution:

$$p_u(d) = (1-b)^2 \cdot f(d) \quad (10)$$

and we can calculate log-likelihood ratio based weights $W(d)$ by integrating (9) and (10) over an interval corresponding to the precision of d (for two observed ages, for example, over $d \pm 1$ years) and taking the logarithm of the ratio of integrals. As in the standard hit-miss model, single or double blanks receive weight 0.

In practice, $f(d)$ must be estimated from training data (often a normal approximation is acceptable) and the probability for a blank b is estimated by the relative frequency of blanks in the entire database. To estimate the other parameters, an EM mixture identifier can be used. The restriction that the four mixture proportions be determined by a_1 and a_2 complicates the maximisation step of the EM algorithm, but can be accounted for in numerical maximisation. For a detailed outline of EM hit-miss mixture identification, see Table 2.

2.1.3 A method to handle correlated record fields

The standard hit-miss model assumes independence between record fields and this allows the total match score for a record pair to be calculated by simple summation of the weights for individual record fields. The independence assumption may, however, lead to over-estimated evidence that two records that match on a set of strongly correlated fields are duplicates, and this may hinder effective duplicate detection.

To reduce the risk for high total match scores driven solely by a group of correlated record fields, we propose a model that accounts for pairwise associations between correlated events. Let $j_1, \ldots, j_m$ denote a set of events related to different fields on the same database record. In the independence model, the probability that these events should co-occur on a record is:

$$P(j_1, \ldots, j_m) = \prod_{t=1}^{m} P(X_t = j_t) = \\ = \prod_{t=1}^{m} (1-b_t)\beta_{j_t} \quad (11)$$

The corresponding total contribution to the match score is:

$$\sum_{t=1}^{m} W_{j_t j_t} = \sum_{t=1}^{m} \log_2\{1 - c_t(1-\beta_{j_t})(1-b_t)^{-2}\} - \sum_{t=1}^{m} \log_2 \beta_{j_t} \quad (12)$$

but this is based on the assumption that the information in the different record fields can be considered independently.

If no assumption of independence can be made, the joint probability for the set of events $j_1, \ldots, j_m$ can only be expressed as:

$$P(j_1, \ldots, j_m) = P(j_1) \cdot P(j_2 \mid j_1) \cdot P(j_3 \mid j_1, j_2) \cdot \\ \cdot \ldots \cdot P(j_m \mid j_1, \ldots, j_{m-1}) \quad (13)$$

However, the amount of data required to reliably estimate $P(j_m \mid j_1, \ldots, j_{m-1})$ increases rapidly with m. As a compromise we propose the following approximation that accounts for pairwise associations only:

$$P(j_1, \ldots, j_m) = P(j_1) \cdot \prod_{t=2}^{m} \max_{s<t} P(j_t \mid j_s) \quad (14)$$

For correlated record fields, (14) may be used instead of (11) to model the joint distribution. Let:

$$j_t^* = \operatorname*{argmax}_{j_s : s<t} P(j_t \mid j_s) \quad (15)$$

$$\beta_{j_t}^* = (1-b_t) \cdot P(j_t \mid j_t^*) \quad (16)$$

Then:

$$W_{jj}^* = \log_2\{1 - c(1-\beta_j^*)(1-b)^{-2}\} - \log_2 \beta_j^* \quad (17)$$

and:

$$\sum_{t=1}^{m} W_{j_t j_t}^* = \sum_{t=1}^{m} \log_2\{1 - c_t(1-\beta_{j_t}^*)(1-b_t)^{-2}\} - \sum_{t=1}^{m} \log_2 \beta_{j_t}^* \\ \approx \sum_{t=1}^{m} \log_2\{1 - c_t(1-\beta_{j_t})(1-b_t)^{-2}\} - \sum_{t=1}^{m} \log_2 \beta_{j_t}^* \\ = \sum_{t=1}^{m} W_{j_t j_t} - \sum_{t=1}^{m} \log_2 \frac{\beta_{j_t}^*}{\beta_{j_t}} \quad (18)$$

Thus, the adjusted match score can be calculated by subtracting a sum of compensating terms from the original match score. Each compensating term can be written on the following form:

$$\log_2 \frac{\beta_{j_t}^*}{\beta_{j_t}} = \log_2 \frac{P(j_t \mid j_t^*)}{P(j_t)} \quad (19)$$

and a shrinkage estimate for this log-ratio has earlier proven useful, as robust strength of association measures to find interesting associations in the WHO drug safety database [1, 15]. This strength of association measure is referred to as the IC and is defined as [1, 15]:

$$IC_{ij} = \log_2 \frac{P(j \mid i)}{P(j)} \quad (20)$$

Shrinkage is achieved through Bayesian inference with a prior distribution designed to moderate the estimated IC values toward the baseline assumption of independence ($IC = 0$) [1, 15]. The advantage of using IC values rather than raw observed-to-expected ratios is that they provide less volatile estimates when little data is available. In order to provide more robust scoring of correlated record fields, we propose IC shrinkage estimates be used to estimate $\log_2 \frac{\beta_{j_t}^*}{\beta_{j_t}}$ in (18).

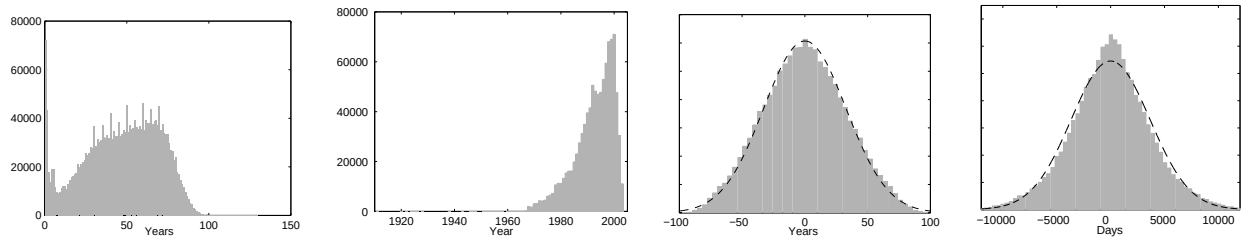
The ordering of events $j_1, \ldots, j_m$ may affect the magnitude of the compensating term in (18) since conditioning is only allowed on preceding events in the sequence. As a less arbitrary choice of ordering, we propose the set be re-arranged in decreasing order of maximal IC value with another event in the set of matched events.

2.2 Fitting a generalised hit-miss model to WHO drug safety data

An adapted hit-miss model was fitted to the WHO drug safety database based on the data available at the end of 2003, including a set of 38 manually identified groups of duplicate records.

Record field	Interpretation	Type	Missing data
DATE	Date of onset	String	23%
OUTCOME	Patient outcome	Discrete (7 values)	22%
AGE	Patient age	Numerical (years old)	19%
GENDER	Patient gender	Discrete (2 values)	8%
DRUGS	Drugs used	14,280 binary events	0.08%
ADRS	ADRs observed	1953 binary events	0.001%
COUNTRY	Reporting country	Discrete (75 values)	0%

Table 3: Record fields used for duplicate detection in the WHO database.



a. Ages b. Dates c. Age differences d. Date differences

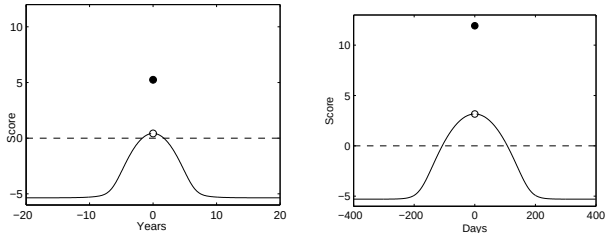
Figure 2: Empirical distributions for ages and dates on records in the WHO database, as well as empirical $f(d)$ functions together with fitted normal distributions.

2.2.1 Implementation

Although the WHO database allows for the transmission and storage of a large amount of data for each individual case report, there are few records that have even the majority of the fields filled in [1]. However, all records in the data set have at least one drug substance, one ADR term and the reporting country listed. For the identification of possible duplicate records, the following record fields were considered the most relevant: date of onset, patient age, patient gender, reporting country, patient outcome, drug substances used and ADR terms observed (drug substances and ADR terms are in fact sets of binary events related to the presence or absence of each). Table 3 lists basic properties for these record fields.

Some data pre-processing was required. Onset dates are related to individual ADR terms, and although there tends to be only one distinct onset date per record, there are 1184 records (0.04% of the database) that have different onset dates for different ADR terms; for those records, the earliest listed onset date was used. For the gender and outcome fields "-" had sometimes been used to denote missing values, and was thus re-encoded as such. Similarly, gender was sometimes listed as N/A which was also considered a missing value. For the age field, a variety of non-standard values were interpreted as missing values and re-encoded as such. Sometimes different age units had been used so in order to harmonise the ages, they were all re-calculated and expressed in years. Observed drug substances are listed as either suspected, interactive or concomitant, but since this subjective judgement is likely to vary between reporters, this information was disregarded.

For large data sets, it is computationally intractable to score all possible record pairs. A common strategy is to group the records into different blocks based on their values for a subset of important record fields and to only score records that are within the same block [9]. For the WHO database, we block records based on drug substances crossed with ADR types so that only record pairs that have at least one drug substance in common and share at least one ADR type (as defined by the System Organ Class, which is a higher level grouping of ADR terms) are scored. In addition to the improvement in computational efficiency, this also reduces the risk for false leads generated by almost identical non-duplicate database records that refer to different reactions in the same patient (see Section 1). While blocking may in theory yield extra false negatives, duplicate records that don't match on at least one drug substance and an ADR type are very unlikely to receive high enough match scores to exceed the threshold for manual review.



a. Age b. Date

Figure 3: Fitted hit-miss mixture model weight functions for age and date, respectively. Note the discrete jump in the weight functions at $d = 0$.

2.2.2 Model fitting

The majority of the hit-miss model parameters are estimated based on the entire data set, but c, a_1 and a_2 rely on the characteristics of identified duplicate records. For the WHO drug safety database there were 38 groups of 2-4 suspected duplicate records available for this purpose. These had been identified earlier by manual review.

Record field	$\hat{a}$	$\hat{b}$	W_{jk}	Maximum W_{jj} value	Minimum W_{jj} value
GENDER	0.051	0.080	-3.22	1.22 (Male)	0.68 (Female)
COUNTRY	0.036	0.000	-3.80	18.45 (Iceland)	1.03 (USA)
OUTCOME	0.101	0.217	-2.05	8.19 (Died unrelated to reaction)	0.97 (Recovered)
DRUGS	0.107	0.001	-2.30	21.23 (non-unique)	4.77 (acetylsalicylic acid)
ADRS	0.387	0.000	-0.68	20.14 (non-unique)	2.77 (rash)

Table 4: Some parameters for the hit-miss model fitted to the WHO database. The W_{jk} value is the weight for a mismatch in that particular record field. The listed W_{jj} values are the maximum and minimum weights for matches on events in that particular record field.

Standard hit-miss models were fitted to the gender, country and outcome record fields. Separate hit-miss models were fitted for individual drug substances and ADR terms, but b and c was estimated for drug substances as a group and for ADR terms as a group (c was estimated based on (7) where $\sum \beta_i^2$ was replaced by the average $\sum \beta_i^2$ for the group). Some of the fitted hit-miss model parameters are displayed in Table 4. As expected, matches on common events such as female gender receive much lower weights than matches on more rare events such as originating in Iceland. The penalty for mismatching ADR terms is significantly lower than that for mismatching drug substances, because discrepancies are more common for ADR terms. This is natural since the categorisation of adverse reactions requires clinical judgement and is more prone to variation.

For the numerical record fields age and date, hit-miss mixture models as described in Section 2.1.2 were fitted. Figure 2 shows empirical distributions in the WHO database for age and date together with the corresponding $f(d)$ functions (note as an aside the digit preference on 0 and 5 for age). Since the empirical $f(d)$ functions for both age and date are approximately normal and since they must be symmetrical by definition ($d = j - i$ and i and j follow the same distribution), we assume normal $f(d)$ functions with mean 0 for both age and date. The variances were estimated by:

$$\hat{\sigma}_2^2 = \frac{\sum_{i=1}^{n} d_i^2}{n} \tag{21}$$

where n is the number of record pairs on which the estimate is based. EM mixture identification as outlined in Table 2 with the estimated values for b and σ_2^2 and with starting values $\hat{a}_1 = 0.1$ and $\hat{a}_2 = 0.1$ yielded the following parameters for the hit-miss mixture model for age:

$$\begin{aligned} &\hat{a}_1 = 0.036 \qquad \hat{a}_2 = 0.010 \qquad \hat{b} = 0.186 \\ &\hat{\sigma}_1 = 2.1 \qquad \hat{\sigma}_2 = 32.9 \end{aligned} \tag{22}$$

and for date:

$$\begin{aligned} &\hat{a}_1 = 0.051 \qquad \hat{a}_2 = 0.010 \qquad \hat{b} = 0.229 \\ &\hat{\sigma}_1 = 50.2 \qquad \hat{\sigma}_2 = 3655 \end{aligned} \tag{23}$$

Because of the limited amount of training data available, we enforced a lower limit of 0.01 for both $\hat{a}_1$ and $\hat{a}_2$. Thus, even though no large deviations in age and date were observed in our training data, the possibility of large errors in these record fields is not ruled out.

A problem with onset date is that quite a large proportion of the records in the data set ($> 15\%$) have incomplete but not altogether missing information (such as 2002-10-? or 1999-?-?). This is straightforwardly taken care of in the hit-miss mixture model by integrating over a wider interval, when calculating the weight. For example, to compare dates 2002-10-? and 2002-10-12, we integrate (9) and (10) from -12 to 20. In practice, this leads to weights around 4.5 for matches on year when information on day and month are missing on one of the records and to weights around 8.0 for matches on year and month when information on day is missing on one of the records.

There tend to be strong correlations between drug substances and ADR terms (groups of drug substances are often co-prescribed and certain drug substances cause certain reactions) so IC based compensation according to Section 2.1.3 was introduced for drug substances and ADR terms as one group.

2.2.3 A match score threshold

Under the hit-miss model, the match score correlates with the probability that two records are duplicates. In order to convert match scores to probabilities, we use a simple form of the mixture model discussed by Belin & Rubin [2]. The assumption is that the match scores for duplicate records follow one normal distribution and the match scores for non-duplicate records follow a different normal distribution. For the WHO database, the empirical match score distributions are approximately normal. We estimated the match score mean and variance for duplicates based on the scores for the 38 duplicates in training data (see Section 2.2.2):

$$\hat{\mu}_{s_2} = 42.96 \qquad \hat{\sigma}_{s_2} = 15.73 \tag{24}$$

and for non-duplicates based on a random sample of 10,000 record pairs:

$$\hat{\mu}_{s_1} = -18.50 \qquad \hat{\sigma}_{s_1} = 8.55 \tag{25}$$

The only relevant data available to estimate the overall proportion of duplicates in the data set was the study of duplicate records in vaccine spontaneous reporting data [14], which found duplication rates around 0.05. Based on $\hat{P}(\text{dup}) = 0.05$ and the estimated match score distributions, we used Bayes formula to compute the probability that a given match score s corresponds to a pair of duplicates:

$$P(\text{dup} \mid s) = \frac{0.05 \cdot \phi(s, \hat{\mu}_{s_2}, \hat{\sigma}_{s_2})}{0.05 \cdot \phi(s, \hat{\mu}_{s_2}, \hat{\sigma}_{s_2}) + 0.95 \cdot \phi(s, \hat{\mu}_{s_1}, \hat{\sigma}_{s_1})} \tag{26}$$

In order to obtain an estimated false discovery rate of below 0.05, the match score threshold for likely duplicates was set at 37.5 since $P(\text{dup} \mid 37.5) = 0.95$ according to (26).

2.2.4 Experimental setup

One experiment was carried out to evaluate the performance of the adapted hit-miss model in identifying the most

Onset date	Age	Gender	Country	Outcome	Drug substances	ADR terms	Score	
?	62	M	USA	Died	3 in total	6 in total	-	
1997-08-??	?	M	USA	Died	3 of 3	3 of 6 + 1	25.19	
1999-06-09	62	M	USA	Died	2 of 3 + 1	2 of 6 + 4	23.66	
1997-09-??	62	M	USA	Died	3 of 3 + 3	2 of 6 + 4	22.92	*
1995-11-29	?	M	USA	Died	2 of 3	3 of 6 + 2	22.82	
1997-08-25	?	M	USA	Died	2 of 3	3 of 6 + 3	22.74	

Table 5: The first difficult template record together with the top 5 records in its list of potential duplicates according to the hit-miss model. The test record is marked with an asterisk.

Onset date	Age	Gender	Country	Outcome	Drug substances	ADR terms	Score	
1997-08-23	40	F	USA	Died	5 in total	4 in total	-	
1997-08-23	40	F	USA	Died	5 of 5	1 of 4 + 4	47.28	
1997-08-23	40	?	USA	Died	4 of 5	2 of 4 + 3	45.75	
1997-08-23	40	?	USA	Unknown	5 of 5	0 of 4 + 4	37.78	
1997-08-??	?	M	USA	Died	3 of 5	3 of 4 + 1	28.52	
?	40	F	USA	Died	3 of 5	3 of 4 + 3	27.09	*

Table 6: The second difficult template record together with the top 5 records in its list of potential duplicates according to the hit-miss model. The test record is marked with an asterisk.

likely duplicates for a given database record. The test data set consisted of the 38 groups of identified duplicates described in Section 2.2.2 and to avoid dependence between training cases, we only used the two most recent records in each group. The most recent record was designated the template record and the second most recent record was designated the test record. In the experiment, each template record was scored against all other records within its block (see Section 2.2.1) in the entire WHO database to see if any other record received a higher match score with the template record than the test record. While the same data set had been used in fitting the hit-miss model, its only impact had been on the proportion of misses in different record fields, so the risk for bias in the performance estimates is slight.

Another experiment was carried out to evaluate the performance of the hit-miss model in discriminating duplicates from random record pairs based on the threshold of 37.5 derived in Section 2.2.1. The test set used in the first experiment could not be used to evaluate the threshold since this data had been used to determine the threshold. However, Norway who is one of the few countries that label duplicate records on submission, had in their last batch in 2004 indicated 19 confirmed duplicates. This allowed for an independent evaluation of the duplicate detection method. Match scores were calculated for all record pairs within the same block (see Section 2.2.1) and those with scores that exceeded the 37.5 threshold were highlighted as likely duplicates.

3. RESULTS

3.1 Duplicate detection for a given database record

The performance at duplicate detection for a given database record was evaluated based on whether or not it was the test record that received the highest match score together with the template record. This was the case for 36 out of the 38 record pairs (94.7%). The two template records for which the test record was not top ranked are listed in Table 5 and Table 6 together with the most likely duplicates as indicated by the hit-miss model. For the first difficult template record, there are no strong matches, and based on a superficial examination, the two top ranked records which are not known duplicates seem as plausible as the test record which is a confirmed duplicate. Thus, while its performance was imperfect for this template record, the hit-miss model's predictions are at least in line with intuition. For the second difficult template record, there are strong matches (match scores ranging from 37.78 to 47.28) with 3 records that are not confirmed duplicates. While these may well be false positive, they could also be undetected duplicates: the records match on most of the fields and although some of the ADR terms differ, a more careful analysis shows that the listed ADR terms relate to liver and gastric problems. Thus, while the hit-miss model failed to identify the known duplicate for this template record, it may have identified 3 that are currently unknown.

3.2 Discriminant duplicate detection

There was a total of 1559 case reports in the last batch from Norway in 2004. The median match score for the 19 known pairs of duplicates was 41.8 and the median match score for all other record pairs (after blocking) was -4.8. Figure 4 displays the match score distributions for the two groups. All in all, 17 record pairs had match scores above 37.5 and out of these, 12 correspond to known duplicates and 5 to other record pairs. Thus, the recall of the algorithm in this experiment was 63% (12 of the 19 confirmed duplicates were highlighted) and the precision was 71% (12 of the 17 highlighted record pairs are confirmed duplicates). However, the threshold of 37.5 was set based on the assumed 5% rate of duplicates in the data set, and following the discussion of precision-recall graphs by Bilenko & Mooney [4] Figure 5 indicates how the precision and the recall varies with different thresholds (an estimated 20% rate of duplicates would give a 35.2 threshold, an estimated 10% rate of duplicates would give a 36.5 threshold and an estimated 1% rate of duplicates would give a 39.6 threshold). To achieve

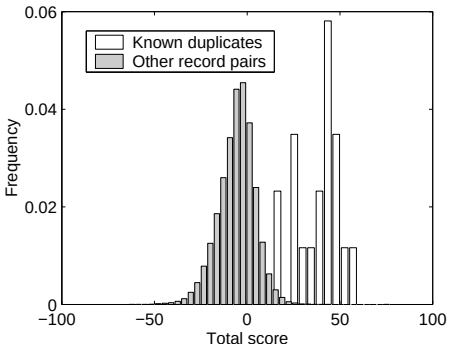


Figure 4: Match score distributions for known duplicates and other record pairs in the Norwegian batch, normalised in order to integrate to 1.

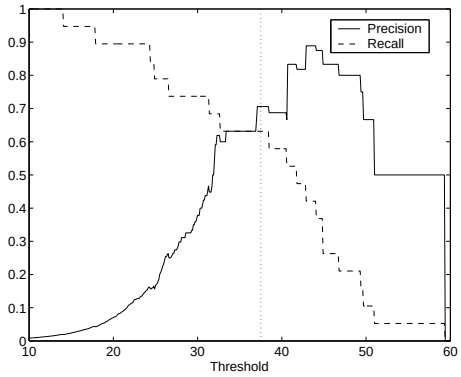


Figure 5: Precision and recall as functions of the threshold, for the discriminant analysis experiment on Norwegian data. The dotted line indicates the selected threshold.

the minimum total number of errors, 11 (2 false positives and 9 false negatives), a threshold between 40.7 and 41.7 must be used. Precision normally tends to 1 as the threshold is increased, but this is not the case in Figure 5, because the highest match score actually corresponds to a pair of records that were not known duplicates. Table 7 lists the three record pairs with highest match scores among record pairs that were not confirmed duplicates and Table 8 lists the three record pairs with lowest match scores among confirmed duplicate record pairs.

3.3 Computational requirements

The experiments were run on a workstation equipped with a 2.2 GHz P4 processor and 1 GB of RAM. Efficient use of the available hardware and optimised data structures reduced computing time and memory requirements so that the initial data extraction and model fitting required a total of 50 minutes. To score a single pair of database records took 6 μs, and to score a database record against the rest of the data set took about 1 second (average block size in the order of 100,000 records). The scoring for all record pairs in the Norwegian data subset (1559 database records), after blocking, took 27 seconds.

4. DISCUSSION

The hit-miss model as implemented on WHO data produced very promising results. For records that are known to have a duplicate, the hit-miss model reliably highlighted the corresponding record (94.7% accuracy). However, only a small proportion of database records have duplicates, so high ranked records are not necessarily duplicates, and in order for the method to be truly effective at duplicate detection, it needed to provide an absolute estimate for the probability that two records are duplicates. The 63% recall and 71% precision in Section 3.2 indicate that the hit-miss model identified the majority of known duplicates, while generating few false leads, which demonstrates its practical usefulness.

The hit-miss model did fail to highlight 7 known duplicates in the Norwegian batch, but from Table 8 it is clear that the amount of information on these records is very scarce: ages, outcomes and onset dates are missing on at least one of the records in each pair and while there are a few matching drug substances and ADR terms, there are at least as many unmatched ones. The lowering of the threshold required to highlight all these duplicates would yield an unmanageable proportion of false leads. We anticipate that any method would require non-anonymised data to be able to identify such duplicates, since lack of data cannot be compensated for with advanced algorithms. This emphasises the need for improved quality of case reports.

Five of the record pairs highlighted in the Norwegian batch were not confirmed duplicates. One of these received the highest match score in the experiment (the top one in Table 7), but did not seem like an obvious pair of duplicates: outcomes are missing, onset dates and ages are close but don't match and none of the registered ADR terms match. On the other hand, 6 out of the 7 drug substances on these two records are the same and this is what generated the unusually high match score. These drug substances are not particularly commonly co-reported (the pairwise associations between them are weak) which further strengthens the evidence. In order to determine the true status of this record pair, we subsequently contacted the Norwegian national centre who confirmed that it was indeed a pair of duplicates: two different physicians at the same hospital had provided separate case reports for the same incident. This demonstrates that the hit-miss model may account for probabilistic aspects of data that are not immediately clear from manual review and that the hit-miss mixture model's treatment of small deviations in numerical record fields may be very useful in practice. The Norwegian centre also provided information on the 4 other record pairs of unknown status that had been highlighted in the study: the record pair with the second highest match score was reported to be a likely but yet unconfirmed duplicate whereas the other three highlighted record pairs were confirmed non-duplicates. However, these case reports had all been provided by the same dentist and all referred to the same drug-ADR combination. Such case reports submitted by the same individual will tend to be similar and difficult to distinguish from true duplicates. With respect to duplicate detection, these record pairs are certainly false leads, but in a different context the detection of such clusters of case reports may be very valuable (since they would generally be considered less strong evidence of a true problem than case reports from independent sources). The Norwegian feedback indicates that the reported 71% precision in Section 3.2 is an under-estimate.

Onset date	Age	Gender	Country	Outcome	Drug substances	ADR terms	Score
2004-04-30	51	F	NOR	?	6 matched, 1 unmatched	0 matched, 3 unmatched	76.97
2004-04-20	50	F	NOR	?			
2003-02-02	57	M	NOR	?	3 matched, 1 unmatched	1 matched, 0 unmatched	42.88
2003-02-02	55	M	NOR	?			
2003-12-16	8	F	NOR	?	1 matched, 0 unmatched	1 matched, 0 unmatched	40.69
2003-12-16	18	F	NOR	?			

Table 7: The three record pairs with highest match scores among record pairs that are not confirmed duplicates in the Norwegian data.

Onset date	Age	Gender	Country	Outcome	Drug substances	ADR terms	Score
?	79	F	NOR	?	1 matched, 0 unmatched	1 matched, 2 unmatched	24.36
?	?	F	NOR	?			
2003-01-07	76	F	NOR	?	1 matched, 1 unmatched	1 matched, 3 unmatched	17.82
?	?	F	NOR	?			
?	43	F	NOR	?	2 matched, 2 unmatched	0 matched, 8 unmatched	14.05
?	?	F	NOR	?			

Table 8: The three record pairs with lowest match scores among non-highlighted confirmed duplicates in the Norwegian data.

The actual precision of the experiment was at least 76% (13/17) and possibly even higher. The reported recall rate may be either under- or over-estimated depending on how many unidentified duplicates remain.

The hit-miss mixture model is a new approach to handling discrepancies in numerical record fields. Like the standard hit-miss model, it is based on a rigorous probability model and provides intuitive weights. For matches, the weights depend on the precision of the matching values: matches on full dates receive weights around 12.0, matches on year and month when day is missing receive weights around 8.0 and matches on year when month and day are missing receive weights around 3.5. Both matches and near-matches are rewarded, and the definition of a near-match is data driven: for the WHO database, age differences within ± 1 year and date differences within ± 107 days receive positive weights and are thus favoured over missing information. There is a limit to how strongly negative the weight for a mismatch will get (see Figure 3), so any large enough deviation is considered equally unlikely. An alternative model for dates which would be useful if typing errors were very common is to model year, month and day of the date as separate discrete variables. The disadvantage of this approach is that absolute differences of just a few days could lead to very negative weights whereas differences of several years may yield positive weights if the two records match on month and day. In the hit-miss model, on the other hand, a pair of dates such as 1999-12-30 and 2000-01-02 contributes +3.18 to the match score, despite the superficial dissimilarity.

The experiments in this article were retrospective in the sense that they evaluated the performance of the algorithms based on what duplicates had already been identified. In the future, we aim to do a prospective study where the hit-miss model is used to highlight suspected duplicates in an unlabelled data subset and follow up the results by manual review. Such a study should allow for more accurate precision estimates and more insight into how the algorithms may be best applied in practice.

The hit-miss model will be used routinely for duplicate detection in in the WHO database. Database wide screens will be carried out regularly and, in addition, duplicate detection can be carried out at data entry and automatically when a case series is selected for clinical review. The rate limiting step in duplicate detection for post-marketing drug safety data is the manual review required to confirm or refute findings, so further testing will be necessary to determine whether the selected threshold is practically useful.

The hit-miss model fitted to the WHO drug safety database in Section 2.2 can be used for duplicate detection in other post-marketing drug safety data sets as well, provided they contain similar information. An alternative approach would be to use the methods described in this paper to fit adapted hit-miss models directly for the data sets of interest, since the properties of different data sets may vary and additional record fields may be available.

5. CONCLUSIONS

In this paper we have introduced two generalisations of the standard hit-miss model and demonstrated the usefulness of the adapted hit-miss model for automated duplicate detection in WHO drug safety data. Our results indicate that the hit-miss model can detect a significant proportion of the duplicates without generating many false leads. Its strong theoretical basis together with the excellent results presented here, should make it a strong candidate for other duplicate detection and record linkage applications.

6. ACKNOWLEDGEMENTS

The authors are indebted to all the national centres who make up the WHO Programme for International Drug Monitoring and contribute case reports to the WHO drug safety database, and in particular to the Norwegian national centre for allowing the evaluation of their data to be used in this paper and for providing rapid assessment of the suspected duplicates. The opinions and conclusions, however, are not necessarily those of the various centres nor of the WHO.

7. 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*, 54:315–321, 1998.

[2] T. Belin and D. Rubin. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90:694–707, 1995.

[3] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM Press, 2003.

[4] M. Bilenko and R. J. Mooney. On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD-2003 workshop on data cleaning, record linkage and object consolidation*, pages 7–12, 2003.

[5] E. A. Bortnichak, R. P. Wise, M. E. Salive, and H. H. Tilson. Proactive safety surveillance. *Pharmacoepidemiology and Drug Safety*, 10:191–196, 2001.

[6] A. D. Brinker and J. Beitz. Spontaneous reports of thrombocytopenia in association with quinine: clinical attributes and timing related to regulatory action. *American Journal of Hematology*, 70:313–317, 2002.

[7] J. Copas and F. Hilton. Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society: Series A*, 153(3):287–320, 1990.

[8] I. R. Edwards. Adverse drug reactions: finding the needle in the haystack. *British Medical Journal*, 315(7107):500, 1997.

[9] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210, 1969.

[10] M. A. Hernandez and S. J. Stolfo. The merge/purge problem for large databases. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 127–138. ACM Press, 1995.

[11] M. Lindquist. Data quality management in pharmacovigilance. *Drug Safety*, 27(12):857–870, 2004.

[12] A. E. Monge and C. Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.

[13] H. B. Newcombe. Record linkage: the design of efficient systems for linking records into individual family histories. *American Journal of Human Genetics*, 19:335–359, 1967.

[14] J. N. Nkanza and W. Walop. Vaccine associated adverse event surveillance (VAEES) and quality assurance. *Drug Safety*, 27:951–952, 2004.

[15] R. Orre, A. Lansner, A. Bate, and M. Lindquist. Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics & Data Analysis*, 34:473–493, 2000.

[16] M. D. Rawlins. Spontaneous reporting of adverse drug reactions. II: Uses. *British Journal of Clinical Pharmacology*, 1(26):7–11, 1988.

[17] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM Press, 2002.

II

Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events

G. Niklas Norén[1,2,*], Andrew Bate[1], Roland Orre[3] and I. Ralph Edwards[1]

[1] *The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden*
[2] *Mathematical statistics, Stockholm University, Stockholm, Sweden*
[3] *NeuroLogic Sweden AB, Stockholm, Sweden*

SUMMARY

Post-marketing drug safety data sets are often massive, and entail problems with heterogeneity and selection bias. Nevertheless, quantitative methods have proven a very useful aid to help clinical experts in screening for previously unknown associations in these data sets. The WHO international drug safety database is the world's largest data set of its kind with over 3 million reports on suspected adverse drug reaction incidents. Since 1998, an exploratory data analysis method has been in routine use to screen for quantitative associations in this data set. This method was originally based on large sample approximations and limited to pairwise associations, but in this article we propose more accurate credibility interval estimates and extend the method to allow for the analysis of more complex quantitative associations. The accuracy of the proposed credibility intervals is evaluated through comparison to precise Monte Carlo simulations. In addition, we propose a Mantel-Haenszel type adjustment to control for suspected confounders.

1. Introduction

Despite great efforts in investigating drug safety before new substances are introduced on the market, some adverse drug reactions (ADR) are not detected until after drug launch. This applies in particular to reactions that have low incidence, occur primarily in groups that tend to be excluded from clinical trials (such as pregnant women or young children), are due to drug interactions or have long times to onset [1]. Screening of spontaneous reports is one of several tools for post-marketing drug safety surveillance [2, 3], and remains the main method for generating hypotheses related to previously unknown adverse drug reactions [4, 5]. In this context, international initiatives have the advantage of accumulating information from all over the world, something which increases the potential for early detection of drug safety problems [6]. At the same time, the massive data sets involved require quantitative methods for efficient knowledge discovery.

The WHO Collaborating Centre for International Drug Monitoring in Uppsala, Sweden (also known as the Uppsala Monitoring Centre or the UMC) holds the world's largest database of

*Correspondence to: Niklas Norén, niklas.noren@who-umc.org

spontaneous reports on suspected adverse drug reaction incidents. The first reports in this data set date back to 1967, and as of November 2004, 75 countries from around the world forward their ADR reports to the UMC. The database currently consists of over 3 million reports, with more than 14,000 distinct drug substances and almost 2,000 distinct ADR terms. Out of the over 20 million possible combinations of one of these drug substances with one of these ADR terms, around 600,000 pairs occur together on at least one report in this data set.

In recent years, several methods for quantitative analysis of spontaneous reporting data sets have been proposed — some Bayesian [7, 9] and others non-Bayesian [11, 12]. Unlike earlier approaches [13, 14], the lack of readily accessible and reliable international usage data has lead these methods to focus on associations within the data sets rather than on proper rates of incidence. Instead of external data, the whole database of reported ADR incidents is thus the reference against which each possible association is compared. Despite the biases in this reference population and despite biases in reporting behaviour and problems with data quality (*e.g.* the highly variable amount of information available on different reports and the phenomenon of duplicate reports), these quantitative methods have proven a useful aid in highlighting drug-ADR combinations for clinical review [15, 16]. One advantage with the study of associations within the data set is that some biases (such as the relative over-reporting of new drug substances) is automatically compensated for [12].

Let dependency derivation denote the screening for quantitative associations between events in a large data set. An obvious difficulty with exploratory data analysis ventures is the multiple comparisons problem: given the large number of possible associations that are evaluated simultaneously, it is hard to attribute a degree of significance to the findings. In addition, for drug safety data sets, even if significant quantitative associations between two events can be identified, they could potentially be driven by confounding variables or reporting biases. As a consequence, we use dependency derivation as a means not to draw final conclusions about possible associations between events in the data set but to generate hypotheses. Any suspicion raised through dependency derivation needs to be further evaluated and tested in some follow-up procedure. In the routine screening of the WHO database for potential drug safety problems, this follow-up procedure consists of clinical evaluation of the individual case reports for each highlighted association, by an international panel of drug safety experts [6].

In screening the database, we are interested in both the estimated strengths of association and the support in data; strong associations are more likely to be indicative of important problems, but with very little support in data even strong quantitative associations are likely to be spurious. The problem with the straightforward use of a test for association is that it may tend to highlight weak associations with large data support [9]. On the other hand, raw strength of association estimates are sensitive to random variation and thus vulnerable to spurious associations. In fact, for these large and sparse data sets, even the use of classical confidence intervals around traditional strength of association measures are insufficient to compensate for a limited data support. As an illustration, consider a drug substance x for which in the first quarter after marketing there were only two case reports in the WHO database. Even for a common ADR term, with say 100,000 reports in total, a single observed report together with this drug by far exceeds the expected number (0.07) and the 95% confidence interval for the log-odds ratio ([0.60 6.14]) excludes 0 and thus indicates a quantitative association. Consequently, if log-odds ratios with confidence intervals were used as the screening criterion, single reports on a particular ADR would suffice to highlight new drugs for clinical review.

Bayesian dependency derivation methods can be used to provide a reasonable balance

between strength of association and support in data [7, 9]. Bayesian inference is sometimes criticised on account of the explicit incorporation of prior assumptions in the analysis, but in dependency derivation this is the greatest advantage over classical inference: the conservative prior distribution (based on the à priori assumption of mutual independence between any two events) moderates the strength of association estimates toward the baseline assumption of no association (especially at low counts) and thereby reduces the risk of highlighting spurious associations.

Since 1998, a Bayesian dependency derivation method has been in routine use to rank quantitative associations between drug substances and ADR terms in the WHO database [7, 8]. The use of this method to highlight drug-ADR combinations for clinical review has been thoroughly tested [15] and integrated into the overall signal detection strategy at the UMC [16, 17, 18, 19]. Several associations first highlighted with this approach have been published in the medical literature [20, 21]. The algorithmic framework used is referred to as the Bayesian Confidence Propagation Neural Network (BCPNN). The BCPNN is a statistical neural network where the nodes correspond to different events and the weights between nodes are proportional to the strength of association between different events. The BCPNN can be used for complex tasks such as classification and unsupervised pattern recognition [22, 23, 24, 25, 26], but for the purpose of dependency derivation, only the weights between nodes in the network (referred to as Information Components or IC values) are of interest. These can be estimated directly from data, so for transparency we shall refer to the use of the BCPNN for Bayesian dependency derivation as IC analysis throughout this article.

In this article, we propose more accurate credibility interval estimates for the prior/posterior distribution of the IC that do not rely on large sample theory. We argue in favour of using the mode as the central IC estimate and show how it can be accurately estimated. In addition, we propose a generalisation of the IC to higher order associations in order to screen for ADR risk factors. We also introduce a Mantel-Haenszel type of adjustment for the IC in order to control for potential confounders in heterogeneous data sets.

2. IC analysis for pairwise dependency derivation

Denote by IC_{xy} the Information Component between events x and y for variables X and Y respectively. The IC is defined as the base 2 logarithm of an observed-to-expected ratio for the joint probability of the two events, where the expected value is calculated under the assumption of mutual independence [7, 8]:

$$IC_{xy} = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

A positive IC value indicates that the two events co-occur more frequently than expected under the assumption of independence, and a negative IC value indicates that they co-occur more rarely. The IC is a function of the unknown probabilities $P(x,y)$, $P(x)$ and $P(y)$, and Bayesian inference is used to estimate the IC value. For convenience, a Dirichlet prior distribution (that is conjugate to the multinomial distribution of data) is used for the probability parameters, since this makes closed form expressions for the posterior distributions of $P(x,y)$, $P(x)$ and $P(y)$ readily available. No such closed form expression is known for the posterior distribution of the IC itself, but in recent work, the use of Monte Carlo simulation based on the closed

form expressions for $P(x, y)$, $P(x)$ and $P(y)$, has been effective in learning more about the shape of the posterior IC distribution [27]. We use this approach to evaluate the accuracy of the approximations proposed in this article.

To simplify the annotation: with respect to the presence or absence of two events x and y, denote by p_{11}, $p_{1\cdot}$ and $p_{\cdot 1}$ the probability parameters for $P(x, y)$, $P(x)$ and $P(y)$, respectively. Similarly, denote by n_{11}, $n_{1\cdot}$ and $n_{\cdot 1}$ the corresponding numbers of observations in the data set. In addition, denote by n_{10} the number of cases where $X = x$ but $Y \neq y$, by n_{01} the number of cases where $X \neq x$ but $Y = y$, and by n_{00} the number of cases where both $X \neq x$ and $Y \neq y$. Denote by p_{10}, p_{01} and p_{00} the corresponding probabilities.

In our data model, the observed counts n_{11}, n_{10}, n_{01} and n_{00} are assumed to follow a $Mn(p_{11}, p_{10}, p_{01}, p_{00}, n_{\cdot\cdot})$ distribution. With a $Di(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})$ prior distribution for p_{11}, p_{10}, p_{01} and p_{00}, it is a standard result from Bayesian statistics that the corresponding posterior distribution is $Di(\gamma_{11}, \gamma_{10}, \gamma_{01}, \gamma_{00})$, where $\gamma_{ij} = \alpha_{ij} + n_{ij}$ (in an abstract sense, the hyper parameters α_{ij} can be thought of as assumed prior observations) [28].

In this model, the marginal distributions of p_{11}, p_{10}, p_{01} and p_{00} are beta. The same is true for $p_{1\cdot} = p_{11} + p_{10}$ and $p_{\cdot 1} = p_{11} + p_{01}$. Specifically:

$$\begin{aligned} p_{11} &\sim Be(\gamma_{11}, \gamma_{10} + \gamma_{01} + \gamma_{00}) \\ p_{1\cdot} &\sim Be(\gamma_{11} + \gamma_{10}, \gamma_{01} + \gamma_{00}) \\ p_{\cdot 1} &\sim Be(\gamma_{11} + \gamma_{01}, \gamma_{10} + \gamma_{00}) \end{aligned} \tag{2}$$

However, since p_{11}, $p_{1\cdot}$ and $p_{\cdot 1}$ are not independent ($p_{1\cdot} = p_{11} + p_{10}$ and $p_{\cdot 1} = p_{11} + p_{01}$), it will sometimes be a coarse approximation to consider the marginal distributions separately as has been done earlier [7, 8], and we will in this article base our analyses on the full Dirichlet distribution.

Some general problems with observed-to-expected ratios should be kept in mind. Observed-to-expected ratios are relevant strength of association measures primarily for events with low expected frequencies where there is virtually no upper limit to the observed-to-expected ratios. In contrast, if the overall frequency of a certain ADR term is as high as, for example, 0.5, the observed-to-expected ratio for its association with a given drug substance can never exceed 2 – even if that ADR term occurs on every report for that drug substance. As a consequence, comparisons between IC values can potentially be misleading if the expected frequencies vary significantly in magnitude. Another problem with observed-to-expected ratios is that there may be a spill-over effect from a large observed number of reports for an event pair to the expected number of reports for that event pair. Specifically, if the drug substance under study is very common and there are unexpectedly many reports on this drug substance with a particular ADR, this may influence the overall prevalence of that ADR term so much that the strength of association is underestimated by the observed-to-expected ratio. These two problems rarely affect pairwise IC analysis between drug substances and ADR terms in standard drug safety data sets much, but may be important in the analysis of other types of events or of smaller data sets. To minimise the risk for misleading results, it may in some situations be sensible to accompany the estimated IC values for highlighted associations with standard log-odds ratios.

2.1. The moderating prior distribution

The aim of IC analysis is to generate useful leads with respect to quantitative associations in a data set. As previously discussed, it is in this context crucial to avoid the highlighting of an

abundance of associations with weak support in data, but at the same time focus on estimated strength of association. With respect to this issue, Bayesian dependency derivation based on a conservative prior distribution has proven instrumental in moderating the estimated strengths of association when data is scarce [7, 9]. The Bayesian moderation in combination with the use of credibility intervals provides an efficient, pragmatic compromise between methods based on statistical significance only (that may be sensitive to weak associations with large data support) and methods based on raw observed-to-expected ratios (that tend to highlight associations with very little data support). Since the impact of the prior distribution diminishes as data accumulates, for combinations with large support there is little difference between Bayesian and classical estimates.

To ascertain moderation of the posterior distribution toward the baseline assumption of independence ($IC = 0$) for all possible associations in all possible data sets, assume a $Di(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})$ prior distribution for p_{11}, p_{10}, p_{01} and p_{00} where:

$$\begin{aligned} \alpha_{11} &= q_{1\cdot} q_{\cdot 1} \cdot \alpha_{\cdot\cdot} \\ \alpha_{10} &= q_{1\cdot} q_{\cdot 0} \cdot \alpha_{\cdot\cdot} \\ \alpha_{01} &= q_{0\cdot} q_{\cdot 1} \cdot \alpha_{\cdot\cdot} \\ \alpha_{00} &= q_{0\cdot} q_{\cdot 0} \cdot \alpha_{\cdot\cdot} \end{aligned} \tag{3}$$

and:

$$\alpha_{\cdot\cdot} = \frac{0.5}{q_{1\cdot} q_{\cdot 1}} \tag{4}$$

and:

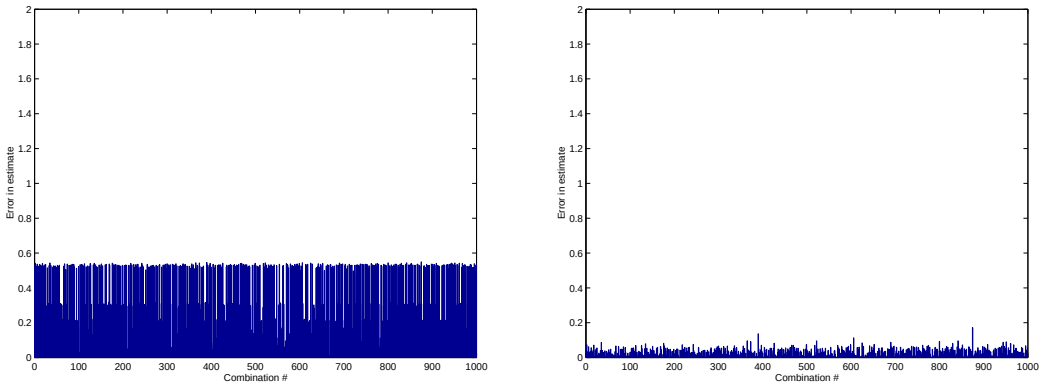
$$\begin{aligned} q_{1\cdot} &= \frac{n_{1\cdot} + 1/2}{n_{\cdot\cdot} + 1} \\ q_{0\cdot} &= \frac{n_{0\cdot} + 1/2}{n_{\cdot\cdot} + 1} \\ q_{\cdot 1} &= \frac{n_{\cdot 1} + 1/2}{n_{\cdot\cdot} + 1} \\ q_{\cdot 0} &= \frac{n_{\cdot 0} + 1/2}{n_{\cdot\cdot} + 1} \end{aligned} \tag{5}$$

This prior distribution incorporates the independence assumption by setting the hyper parameters proportional to the products of the corresponding marginal probabilities (in fact to posterior mean estimates for the marginal probabilities based on $Be(1/2, 1/2)$ hyper priors). The benefit of this is that the IC_{map} always lies between 0 and the raw observed-to-expected log-ratio and that:

$$\begin{aligned} \lim_{n_{1\cdot}, n_{\cdot 1} \to 0} IC_{map} &\approx 0 \\ \lim_{n_{\cdot\cdot} \to 0} IC_{map} &= 0 \end{aligned} \tag{6}$$

which is important for computational stability.

In the abstract sense mentioned above, the moderating prior distribution is equivalent to an assumed extra batch of data where the two events under study are independent, co-occur 0.5



a. Posterior mean estimate b. Maximum à posteriori estimate

Figure 1. Errors in posterior mean and maximum à posteriori estimates for the IC distribution. This figure shows deviations from the Monte Carlo simulated values of estimates based on (7) for the mean and the mode, respectively. 1,000 randomly selected drug-ADR pairs in the WHO database were used and each Monte Carlo simulation was based on 50,000 draws.

times and where the marginal probabilities for the two events are approximately the same as in the real data set. With this approach, the prior sample size $\alpha_{..}$ may vary, but α_{11} always equals 0.5 and since it is primarily α_{11} that determines the shape of the IC distribution [27], the shape of the prior distribution will be approximately the same for all associations under study.

This prior distribution is based on the same principles as the prior previously used in IC analysis (see [7] or [8]). The major differences are that the new prior is based on the joint Dirichlet distribution for the model parameters (instead of independent beta distributions) and that the prior sample size has been halved. The reduction in prior sample size yields a more diffuse prior distribution that better reflects our initial uncertainty about the IC values in the WHO database. For data sets with different characteristics (size, sparsity, heterogeneity) than the WHO database, the factor 0.5 in the expression for $\alpha_{..}$ should be adjusted. It may, for example, be sensible to reduce this factor (and thus the moderating effect of the prior distribution) for smaller data sets.

2.2. Central IC estimates

Arbitrarily accurate estimates for the posterior mean (p.m.e.) of the IC distribution are available [29]. However, as the IC distribution is generally unimodal, maximum á posteriori (m.a.p.) estimates may be used for central estimates instead. The main advantage of the m.a.p. estimate is that it is well suited for use in stratified IC analysis (see Section 2.4) and that it has the intuitive property of being equal to 0 when the estimated joint probability equals the product of the estimated marginal probabilities. In addition, the concept of a most likely value for an unknown parameter is perhaps more natural than that of an expected value, and this is an important aspect in the drug safety application, where the results must be interpretable for non-statisticians.

	$n_{11}=1$	$n_{11}=2$	$n_{11}=3$	$n_{11}=4$	$n_{11}=5$
IC_{pme}	0.53	0.30	0.21	0.16	0.13
IC_{map}	0.04	0.02	0.01	0.01	0.01

Table I. Average error for (7) as estimate of the IC mean and mode, respectively, at different values for the joint count n_{11}.

We propose the the following m.a.p. estimate:

$$IC_{map} \approx \log_2 \frac{E[p_{11}]}{E[p_{1\cdot}]E[p_{\cdot 1}]} \tag{7}$$

The same expression has been used earlier as a crude estimate for the IC mean [7, 8]. To study the accuracy of this expression as an estimate for on one hand the mean and on the other hand the mode of the IC distribution, estimated values were compared to Monte Carlo simulated values based on 50,000 draws from each posterior IC distribution (the mode of the simulated IC distributions was estimated based on the empirical relationship $mode \approx 3 \cdot median - 2 \cdot mean$ for unimodal curves of moderate asymmetry [30]). A random subset of 1,000 drug-ADR combinations that occur in the WHO database were used for evaluation. Throughout, the moderating prior distribution described in Section 2.1 was used. The results are displayed in Figure 1. Clearly, (7) is a better estimate of the mode than of the mean.

2.3. Improved IC credibility interval estimates

Denote by IC_{025}, the 2.5 percentile of the posterior IC distribution. This is the lower limit of a two-sided 95% credibility interval for the IC, by which associations are typically ranked in IC analysis [7, 8]. The use of a lower credibility interval limit accounts for uncertainty in a conservative manner. The idea is to choose an estimate so that the true value is greater than the estimate with a given degree of certainty (here 97.5%). Together with the moderating prior distribution (see Section 2.1) this helps to reduce the number of false leads generated by IC analysis.

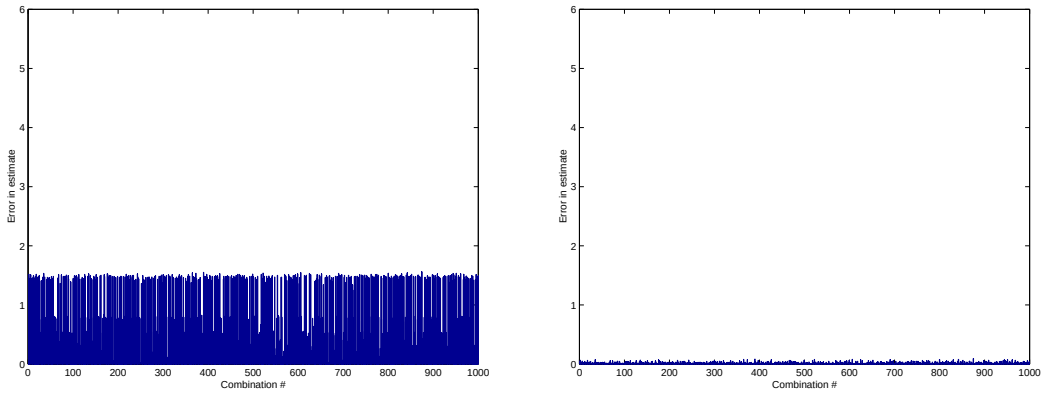
The IC credibility interval estimates were previously based on a normal approximation for the IC distribution [7, 8]. Monte Carlo experiments indicate that while the IC distribution tends to a normal distribution asymptotically (for large n_{11}), the assumption leads to a rather crude approximation for rare pairs of events ($n_{11} \leq 10$). Since more than 80% of the observed drug-ADR pairs in the WHO database fall into this critical category, the need for improvement is clear [27]. The use of brute force Monte Carlo simulation to estimate the posterior percentiles would give arbitrarily accurate estimates, but at too high a cost in computational complexity. Instead, we propose an approach based on an approximate formula for the difference between the mode and the lower credibility interval limit for the IC distribution.

Let Δ_{025} denote the true difference between IC_{map} and IC_{025}. Given estimates for IC_{map} and Δ_{025}, IC_{025} can be estimated as follows:

$$\hat{IC}_{025} = \hat{IC}_{map} - \hat{\Delta}_{025} \tag{8}$$

Empirical testing suggests that functions of the following general form model Δ_{025} well (A_r and B_r are fitted parameters):

$$\Delta_{025}(\gamma_{11}) = A_r \cdot \gamma_{11}^{-1/2} + B_r \cdot \gamma_{11}^{-3/2} \tag{9}$$



a. Normal approximation b. Proposed approximation

Figure 2. Errors in IC_{025} estimates. This figure shows deviations from Monte Carlo simulated values of the two IC_{025} estimates for 1,000 randomly selected drug-ADR pairs in the WHO database. 50,000 draws were used in each Monte Carlo simulation.

	$n_{11}=1$	$n_{11}=2$	$n_{11}=3$	$n_{11}=4$	$n_{11}=5$
Normal approximation	1.47	0.78	0.52	0.38	0.31
Proposed approximation	0.06	0.07	0.06	0.04	0.04

Table II. Average error in the IC_{025} estimates for different values of the joint count n_{11} among the 1,000 drug-ADR pairs from the WHO database.

In particular:

$$\lim_{\gamma_{11}\to\infty} \Delta_{025}(\gamma_{11}) = 0 \tag{10}$$

and because there are only 2 fitted parameters, there is little risk for over-fitting.

By using different parameters A_r and B_r for different ratios $r = \gamma_{11}/\min(\gamma_{1\cdot}, \gamma_{\cdot 1})$, the impact of the smaller of the two marginal parameters may be accounted for. We have estimated constants A_r and B_r for 11 different values of r $(0.0, 0.1, \ldots, 1.0)$, and use linear interpolation in between. Thus, for a given ratio r, the value of Δ_{025} is estimated by the weighted average of the Δ_{025} values for the two closest values of r for which there are fitted constants A_r and B_r available. For details about fitting the parameters in (9) and their values for different r, see Appendix I.1.

To evaluate the accuracy of the proposed approach to estimate IC_{025}, we compared Monte Carlo simulated values based on 50,000 draws to the estimated values, for the same data set as in Section 2.2. For comparison, the accuracy of the normal approximation [7, 8] was also evaluated. The results are displayed in Figure 2. Clearly, the proposed approximation is more accurate.

2.4. Stratified IC analysis

Although the purpose of dependency derivation is hypothesis generation, and a certain number of false leads is acceptable in this context, it is important to keep the proportion of false

Stratum	$n_{polio,sids}$	$n_{polio\cdot}$	$n_{\cdot sids}$	$n_{..}$	$IC_{polio,sids}$
unspecified	25	1126	87	572573	5.25 (4.64)
0 - 1 month	29	1408	79	9066	1.21 (0.73)
2 months - 4 years	203	30068	508	155209	1.04 (0.87)
5 - 11 years	0	5232	3	80140	-0.48 (-11.10)
12 - 16 years	0	299	0	63911	0.00 (-10.65)
17 - 69 years	0	461	13	1669422	-0.01 (-10.67)
70+ years	0	10	0	453481	0.00 (-10.66)

Table III. Stratum specific IC values for the association between SIDS and the Polio virus vaccine in different age groups. The numbers listed in the rightmost column are IC_{map} estimates with the moderating prior (IC_{025} estimates in brackets).

leads at a minimum. One approach to improving the specificity is to detect and control for potential confounders. As for other epidemiological applications, adjusted overall estimates may be quoted when there is no suspicion of effect modification, otherwise stratum specific estimates should be used [31]. There are at least two different ways to adjust the IC for potential confounders.

The most obvious adjusted IC estimate is a weighted average of the stratum specific IC values as previously suggested [32]. However, this approach requires a careful selection of stratification variables, because it is particularly sensitive to data thinning. Strata with few or no observations of the event combination of interest will yield unreliable stratum specific IC estimates, and since the weights in calculating the adjusted estimate are not necessarily correlated to the reliability of the estimates, this may lead to an unreliable adjusted IC estimate. Indeed, tentative experiments based on Monte Carlo simulation indicate that this approach to adjusting the IC typically leads to wider credibility intervals than for the unadjusted IC, which, in addition to the loss in precision, is a technical disadvantage since it makes more difficult the derivation of accurate credibility intervals.

An alternative approach to calculating adjusted IC estimates is to use a Mantel-Haenszel type of adjustment where the denominator in the IC_{map} formula is equal to the weighted average of the expected joint probabilities in the different strata:

$$IC_{map} \approx \log_2 \frac{E[p_{11}]}{\sum_{k=1}^{n} E[p_{1\cdot|k}]E[p_{\cdot 1|k}] \cdot E[p_{\cdot\cdot k}]} \tag{11}$$

Since the numerator is not affected by the adjustment, the spread of the adjusted IC can be expected to be similar to that for the unadjusted IC, and empirical testing supports this assumption. Consequently, the approximate credibility intervals proposed in Section 2.3 may be used for the adjusted IC_{025} as well.

As an illustration of the general usefulness of stratified IC analysis, we have investigated the association in the WHO database between the terms *sudden infant death syndrome* and *Polio virus vaccine live oral*. The unadjusted IC_{map} estimate for this association is 4.78 and the corresponding IC_{025} estimate is 4.63. However, since the Polio virus vaccine is typically given to small children and only small children suffer from SIDS (the sudden infant death syndrome), this is likely to be confounded by age [9].

There are 7 predefined age groups in the WHO database: *unspecified*, *0 - 1 month*, *2 months*

IC_{025}	Old +	Old -
New +	80363	616
New -	3532	522707

Table IV. A cross-classification of the observed drug-ADR pairs in the WHO database, with respect to the signs of the IC_{025} values for the two methods.

- 4 years, *5 - 11 years*, *12 - 16 years*, *17 - 69 years* and *70+ years*. Table III displays stratum specific IC values for the association between SIDS and the Polio virus vaccine for these age groups. Based on this stratification, the adjusted IC_{map} estimate according to (11), is 1.19 and the corresponding IC_{025} estimate is 1.00. Clearly the stratification by age reduces the apparent strength of association. At the same time, the relatively strong association between SIDS and the Polio virus vaccine in the *age: unspecified* stratum renders dubious the listing of any overall IC estimate (adjusted or not). In this situation, a list of stratum specific IC values is probably a more appropriate output. Please note that a proper examination of this quantitative association would require the consideration of other potential confounders as well.

Some problems with routine stratification by a limited set of predefined variables have been pointed out previously [33]. For the WHO database, we use association specific stratification in the post-processing of clinically interesting drug-ADR pairs.

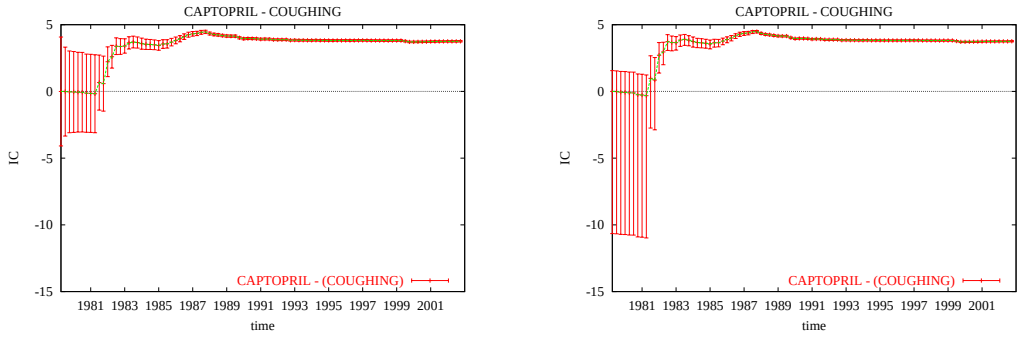
2.5. Example: a scan for drug-ADR associations in the WHO database

To study in practice, the impact of the proposed changes to IC analysis (new prior distribution and improved credibility interval estimates). We have carried out a complete scan of the WHO database (as of quarter 3, 2003) with both methods.

Table IV displays a cross-classification of all observed drug-ADR pairs in the WHO database with respect to whether the IC_{025} values are positive or negative (this is the threshold used in routine screening of the WHO ADR database) with the old and the new approach respectively. Clearly, the agreement between the two approaches is quite good: with respect to this threshold, the two methods differ for only around 4,000 out of the close to 600,000 observed drug-ADR pairs and Cohen's kappa measure is 0.97 (a Cohen's kappa of 1 would indicate perfect agreement). Where the two methods differ, the new approach seems to be somewhat more conservative, but there are event pairs for which the new but not the old IC_{025} estimate exceeds 0. These tend to have low joint counts n_{11} (ranging from 3 to 12) and low marginal counts $n_{1.}$ for the drug (ranging from 3 to 68 in all but three cases, for which the values are significantly higher). In particular, the new approach alone highlights 84 event pairs where there is 3 reports in total for the drug – all on the same reaction. Because these event pairs may correspond to important problems for recently marketed drugs, it is a strength from a monitoring perspective that the new approach highlights them.

2.6. Example: a captopril-coughing time scan

To further examine the practical impact of the proposed changes to IC analysis, we studied the evolution in time of the IC between the drug substance captopril and the ADR term coughing with the two approaches. The association between captopril and coughing has been well known since 1986, but earlier work has shown that if IC analysis had been in use at the



a. Old approach b. New approach

Figure 3. The evolution over time of the IC between captopril and coughing, with the old and the new approach. Central estimates together with 95% credibility intervals are marked in the plot.

time a quantitative association between captopril and coughing would have been highlighted already in 1981 [7]. We were interested to see whether the proposed changes to IC analysis would delay or expedite the highlighting of this quantitative association in the database.

Figure 3 displays the change over time for the IC central estimates (together with 95% credibility interval estimates) for the captopril-coughing quantitative association, based on the old and new IC analysis approaches. The credibility interval estimates differ when the joint count is low, but the association would be highlighted in the same quarter regardless of which approach was used.

3. IC analysis for higher order dependency derivation

The IC as defined in Section 2 is a strength of association measure for pairs of events only, but there is often an interest in higher order associations. In the drug safety application, this may include drug-drug interactions or three way associations involving a drug substance, an ADR term and another risk factor (*e.g.* age or gender). Generally, a higher order strength of association measure should capture disproportionality in the occurrence of groups of events in the data set, which is not explicable by lower order associations. For three way associations, we would be interested in sets of three events that occur unexpectedly often even when pairwise associations between the events are accounted for.

The usefulness of extending the IC value to higher order associations is not necessarily limited to the dependency derivation application. Higher order IC values could be introduced in both the feedforward and the recurrent BCPNN in order to improve performance in classification and unsupervised pattern recognition, respectively.

An extension of the IC to third order associations was previously proposed in [8], but this did not compensate for pairwise associations. We, instead, propose the following definition of the third order IC:

$$IC_{xyz} = IC_{xy|z} - IC_{xy} \tag{12}$$

where:

$$IC_{xy|z} = \log_2 \frac{P(x, y \mid z)}{P(x \mid z)P(y \mid z)} \tag{13}$$

The logic behind this definition of the third order IC is that if there is a positive third order association, the presence of the third event should make the pairwise association between the other two events stronger (*i.e.* $IC_{xy|z}$ should exceed IC_{xy} etc). Conversely, if there is a negative third order associations, the presence of the third event should make the pairwise association between the other two events weaker.

It is easy to show that this definition is symmetric in x, y and z:

$$\begin{aligned} IC_{xyz} &= IC_{xy|z} - IC_{xy} = \\ &= IC_{xz|y} - IC_{xz} = \\ &= IC_{yz|x} - IC_{yz} \end{aligned} \tag{14}$$

since with simple algebraic operations, we can re-express IC_{xyz} as:

$$\begin{aligned} IC_{xyz} &= \log_2 \frac{P(y, z \mid x)}{P(y \mid x)P(z \mid x)} - \log_2 \frac{P(y)P(z)}{P(y, z)} = \\ &= \log_2 \frac{P(x, y, z)P(x)P(y)P(z)}{P(x, y)P(x, z)P(y, z)} \end{aligned} \tag{15}$$

The third order IC can be seen as an observed-to-expected ratio, where the expected value accounts for both main effects and pairwise interactions. To see this, let:

$$W_{x_1 \ldots x_n} = \frac{P(x_1, \ldots, x_n)}{P(x_1) \cdot \ldots \cdot P(x_n)}$$

Then the third order IC_{xyz} may be re-expressed as:

$$IC_{xyz} = \log_2 \frac{P(x, y, z)}{P(x)P(y)P(z)W_{xy}W_{xz}W_{yz}} \tag{16}$$

which is an approximate observed to expected ratio accounting for pairwise associations as well as marginal probabilities.

The generalisation of the IC to even higher orders is straightforward. For example, the fourth order IC can be defined as follows:

$$\begin{aligned} IC_{xyzv} &= IC_{xyz|v} - IC_{xyz} = IC_{xyv|z} - IC_{xyv} = \\ &= IC_{xvz|y} - IC_{xvz} = IC_{yzv|x} - IC_{yzv} \end{aligned} \tag{17}$$

which gives:

$$IC_{xyzv} = \ldots = \log_2 \frac{P(x, y, z, v)}{P(x)P(y)P(z)W_{xy}W_{xz}W_{xv}W_{yz}W_{yv}W_{zv}W_{xyz}W_{xyv}W_{xzv}W_{yzv}} \tag{18}$$

As desired, the approximate expected joint probability in the denominator accounts for both second and third order associations in addition to the marginal probabilities.

Most of the theory developed in Section 2 for pairwise IC values holds approximately for higher order IC values. A third order IC m.a.p. estimate similar to that for pairwise IC is:

$$IC_{map} \approx \log_2 \frac{E[p_{111}]E[p_{1\cdot\cdot}]E[p_{\cdot 1\cdot}]E[p_{\cdot\cdot 1}]}{E[p_{11\cdot}]E[p_{1\cdot 1}]E[p_{\cdot 11}]} \tag{19}$$

	n_{xyz}	$n_{xy\cdot}$	$n_{x\cdot z}$	$n_{\cdot yz}$	$n_{x\cdot\cdot}$	$n_{\cdot y\cdot}$	$n_{\cdot\cdot z}$	$n_{\cdots}$	IC_{xyz}
Drug: ketoconazole	5	63	27	11	6083	3695	5071	3176114	2.32 (1.08)
Age: 17-69 years	52	63	3764	2046	6083	3695	1756414	3176114	0.41 (0.20)
Gender: female	38	63	3427	1607	6083	3695	1753445	3176114	0.43 (0.09)
Country: USA	45	63	2718	2342	6083	3695	1478959	3176114	0.26 (-0.05)
Country: Germany	8	63	174	469	6083	3695	195102	3176114	0.94 (-0.18)

Table V. The top 5 third order IC values with terfenadine and ventricular fibrillation. The numbers listed are m.a.p. estimates with the moderating prior distribution (IC_{025} estimates in brackets).

Credibility intervals for third order IC values may be calculated with the formula proposed in Section 2.3 if, in the definition of r, $\min(\gamma_{1\cdot}, \gamma_{\cdot 1})$ is replaced by $\min(\gamma_{1\cdot\cdot}, \gamma_{\cdot 1\cdot}, \gamma_{\cdot\cdot 1})$. Adjustment of higher order IC values to control for confounders is also possible. For third order IC values, the Mantel-Haenszel adjusted m.a.p. estimate is:

$$IC_{map} \approx \log_2 \frac{E[p_{111}]}{\sum_{k=1}^{n} \frac{E[p_{11\cdot|k}]E[p_{1\cdot 1|k}]E[p_{\cdot 11|k}]}{E[p_{1\cdot\cdot|k}]E[p_{\cdot 1\cdot|k}]E[p_{\cdot\cdot 1|k}]} \cdot E[p_{\cdot\cdot k}]} \tag{20}$$

Furthermore, it is straightforward to generalise the moderating prior distribution described in Section 2.1 to higher order IC values (see Appendix I.2).

3.1. Example: A risk factors scan

Higher order IC analysis may be used to search for factors that influence the risk of a certain ADR given a particular drug. If, for example, the third order IC between a certain drug substance x, a certain ADR term y and a certain age group z were positive, this may indicate that patients of age group z are more prone to experiencing x-induced y than the population in general. Routine scans for third order IC values between a drug substance, an ADR term and some other factor (*e.g.* a certain gender or an age groups) may therefore be used to generate hypotheses with respect to potential high risk groups of patients. Positive higher order IC values may also be indicative of confounding, but for confounders, further investigation will show no significant variation in the IC values over the different strata.

Terfenadine was withdrawn due to concerns about its cardiotoxicity. Additionally, terfenadine and ketoconazole are known to interact so that the risk of heart problems is higher when the two are co-administered. Indeed, there are 5 reports on terfenadine, ketoconazole and ventricular fibrillation in the WHO database and the corresponding third order IC value is 2.32 with a lower credibility interval limit of 1.08. If we were to examine all three way associations between terfenadine and ventricular fibrillation and other events related to age, country, gender or other medication, there are 27 other events that occur at least once together with terfenadine and ventricular fibrillation on reports in the data set. Out of these, only 2 events other than the co-administration of ketoconazole have positive third order IC_{025} values with terfenadine and ventricular fibrillation (see Table V for the top 5 associations with respect to IC_{025} values). Based on this analysis, ketoconazole is clearly the most influential risk factor for this association.

4. Discussion

The analysis of spontaneous reporting data remains the cornerstone of post-marketing drug safety surveillance. Despite problems with data heterogeneity, it is the most important source of information for discovering previously unknown adverse effects from drugs after they are introduced on the market. *IC* analysis has proven to be an efficient method for exploratory quantitative analysis of post-marketing drug safety data [15] that while meeting the computational requirements also provides sophisticated protection against spurious associations. However, *IC* analysis as originally implemented [7, 8] is based on large sample approximations, and despite the large total number of reports in the WHO drug safety database the number of reports on a given drug-ADR pair is typically small (due to the large number of drug substances and ADR terms involved). Thus there is a clear need for the improved credibility intervals proposed in this article, and the results presented in Section 2.3 indicate that they do lead to improved accuracy and may allow for earlier discovery of problems related to recently marketed drug substances. These results are based on randomly selected drug-ADR pairs from the WHO database, but we expect the conclusions to hold generally for rare events in large and sparse data sets.

The Mantel-Haenszel adjustment for the *IC* proposed in Section 2.4 is important in that it will allow for robust exploratory data analysis in the presence of confounding. However, more research is needed to specify efficient strategies for how and when to carry out stratified analyses of spontaneous reporting data. It is, at present time, unclear whether routine adjustment by set of pre-defined variables for all event pairs in the database is to be preferred over unadjusted estimates in the initial screening of the database [33]. If higher order *IC* analysis or other sophisticated pattern recognition methods could be used for automated confounder detection, this may allow for data driven association specific adjustment by suspected confounders, and we aim to investigate this further in the future. The strong association between SIDS and the Polio vaccine in the *age unspecified* stratum of the WHO ADR database (see Section 2.4) is likely to be due to residual confounding and emphasises the problem of missing data for the stratification variables. This issue too needs to be resolved before optimal use of stratified dependency derivation is possible.

While the quantitative improvements for pairwise *IC* analysis proposed in Section 2 are refinements of the existing methodology, the generalisation to higher order associations in Section 3 allows for altogether new types of analysis related to complex quantitative associations. In combination with our methods for unsupervised pattern recognition [25], the methods presented in this article provide a comprehensive range of techniques for efficient knowledge discovery in spontaneous reporting data. An alternative approach to studying higher order associations would be to fit a generalised linear model with interaction terms, and in a similar spirit, other groups have proposed observed-to-expected ratios where the expected frequency is calculated based on a fitted log-linear model [10]. The advantage of higher order *IC* analysis in this context is that it is more direct (it does not require iterative methods for fitting) and allows for local analysis (in the sense that the higher order *IC* value for a certain set of events is only influenced by the joint and marginal counts for that specific set of events). Drug-drug interaction detection is a type of higher order association which is particularly important in the quantitative analysis of spontaneous reporting data and several approaches have been proposed [34, 35, 10]. In theory there is no obvious reason why higher order *IC* analysis could not be used to screen for drug interactions as well as any other risk factors,

but there has recently been a tendency to focus on more simple methods for the detection of drug-drug interactions [36], which indicates that more research into the basic characteristics of drug-drug interactions spontaneous reporting may be needed to resolve this issue successfully.

5. Conclusions

Earlier, IC analysis has proven useful in hypothesis generation with respect to quantitative associations in large drug safety data sets. In this article we have proposed improved methods for posterior inference in IC analysis, including an accurate estimate for the mode and significantly improved credibility interval estimates. In addition, we have extended the IC strength of association measure to higher order associations and illustrated the usefulness of this on real world data. An adjustment of the IC to control for potential confounders has also been described and applied to real world data.

REFERENCES

1. Evans SJ. Pharmacovigilance: a science or fielding emergencies? *Statistics in Medicine*, 2000. **19**(23):3199–3209.
2. Edwards IR. Spontaneous reporting–of what? Clinical concerns about drugs. *British Journal of Clinical Pharmacology* 1999; **48**(2):138–41.
3. Edwards IR. Spontaneous ADR reporting and drug safety signal induction in perspective. To honour Professor Jens Schou. *Pharmacology & Toxicology* 2000; **86**(s1):16–19.
4. Rawlins MD. Spontaneous reporting of adverse drug reactions. I: the data. *British Journal of Clinical Pharmacology* 1988; **26**(1):1–5.
5. Rawlins MD. Spontaneous reporting of adverse drug reactions. II: Uses. *British Journal of Clinical Pharmacology* 1988; **26**(1):7–11.
6. Edwards IR, Olsson S. WHO Programme - global monitoring. In *Pharmacovigilance*, Mann RD, Andrews EB (eds). Wiley:Chichester, 2002; 169–182.
7. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. *European Journal for Clinical Pharmacology* 1998; **54**:315–321.
8. Orre R, Lansner A, Bate A, Lindquist M. Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics & Data Analysis* 2000; **34**:473–493.
9. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting systems. *American Statistician*, 1999; **54**:177–202.
10. DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. In *Knowledge Discovery and Data Mining*, 2001; 67–76.
11. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 2001. **10**(6):483–486.
12. van der Heijden PG, van Puijenbroek EP, van Buuren S, van der Hofstede JW. On the assessment of adverse drug reactions from spontaneous reporting systems: the influence of under-reporting on odds ratios. *Statistics in Medicine*, 2002; **21**(14):2027–2044.
13. Norwood PK, Sampson AR. A statistical methodology for postmarketing surveillance of adverse drug reaction reports. *Statistics in Medicine*, 1988. **7**(10):1023–1030.
14. Praus M, Schindel F, Fescharek R, Schwarz S. Alert systems for post-marketing surveillance of adverse drug reactions. *Statistics in Medicine*, 1993. **12**(24):2383–2393.
15. Lindquist M, Ståhl M, Bate A, Edwards IR and Meyboom RH. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. *Drug Safety* 2000; **23**(6):533–542.
16. Lindquist M, Edwards IR, Bate A, Fucik H, Nunes AM, Ståhl M. From association to alert - a revised approach to international signal analysis. *Pharmacoepidemiology and Drug Safety* 1999; **8**:15–25.
17. Bate A, Lindquist M, Orre R, Edwards IR, Meyboom RH. Data-mining analyses of pharmacovigilance signals in relation to relevant comparison drugs. *European Journal for Clinical Pharmacology*, 2002. **58**(7):483–490.

18. Ståhl M, Lindquist M, Edwards IR, Brown EG. Introducing triage logic as a new strategy for the detection of signals in the WHO drug monitoring database. *Drug Safety*, In Press.
19. Ståhl M, Edwards IR, Bowring G, Kiuru A, Lindquist M. The usefulness and use of signals from the WHO database by national pharmacovigilance centres - results from a questionnaire. *Drug Safety*, In Press.
20. Coulter DM, Bate A, Meyboom RH, Lindquist M, Edwards IR. Antipsychotic drugs and heart muscle disorder in international pharmacovigilance: data mining study. *British Medical Journal* 2001. **322**(7296): 1207–1209.
21. Sanz EJ, De-las-Cuevas C, Kiuru A, Bate A, Edwards IR. Selective serotonin reuptake inhibitors in pregnant women and neonatal withdrawal syndrome: a database analysis *The Lancet* 2005. **365**: 482–487.
22. Lansner A, Ekeberg Ö. A one-layer feedback artificial neural network with a Bayesian learning rule. *International Journal of Neural Systems* 1989; **1**:77–87.
23. Holst A, Lansner A. A higher order Bayesian neural network for classification and diagnosis. In *Applied Decision Technologies: Computational Learning and Probabilistic Reasoning*; Gammerman A (ed). Wiley: New York, 1996; 251–260.
24. Orre R, Lansner A. Pulp quality modelling using Bayesian mixture density neural networks. *Journal of Systems Engineering* 1996; **6**:128–136.
25. Orre R, Bate A, Norén GN, Swahn E, Arnborg S, Edwards IR. A Bayesian recurrent neural network for unsupervised pattern recognition in large incomplete data sets. *International Journal of Neural Systems* 2005; **15**(3):207–222.
26. Norén GN, Orre R. Case based imprecision estimates for Bayes classifiers with the Bayesian bootstrap. *Machine Learning* 2005; **58**:79–94.
27. Norén N. A Monte Carlo method for Bayesian dependency derivation. Master's thesis, Chalmers University of Technology, 2002.
28. Gelman A, Carlin JB, Stern HS, Rubin DB *Bayesian Data Analysis* (1st edn). Chapman & Hall: 1995.
29. Koski T, Orre R. Statistics of the Information Component in Bayesian neural networks. Technical report, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, 1998.
30. Kenney JF, Keeping ES *Mathematics of Statistics, Pt 1* (3rd edn). Van Nostrand: 1962; 50–54.
31. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 1959; **22**:719–748.
32. Gould L. Practical pharmacovigilance analysis strategies. *Pharmacoepidemiology and Drug Safety*, 2003; **12**:559–574.
33. Bate A, Edwards IR, Lindquist M, Orre R. Violation of Homogeneity: A Methodological Issue in the Use of Data Mining Tools. The authors' reply. *Drug Safety*, 2003; **26**(5):363–366.
34. Amery WK. Post-marketing drug safety management: a pharmaceutical industry perspective. *International Journal of Risk and Safety in Medicine*, 1994; **5**:67–270.
35. van Puijenbroek EP, Egberts ACG, Meyboom RHB, Leufkens HGM. Signalling possible drug-drug interactions in a spontaneous reporting system: delay of withdrawal bleeding during concomitant use of oral contraceptives and itraconazole. *British Journal of Clinical Pharmacology*, 1999; **47**:689–693.
36. Yang X, Fram DM. Using disproportional analysis as a tool to explore severe drug-drug interactions in AERS database. *Pharmacoepidemiology and Drug Safety*, 2004; **13**(1):S247.

APPENDIX

I.1. Δ_{025} *parameter fitting*

Constants A_r and B_r for 11 different values of r $(0.0, 0.1, \ldots, 0.9, 1.0)$ were fitted to Equation 9 based on simulated Δ_{025} values where γ_{11} ranged from 1 to 100, $\gamma_{1\cdot} = \gamma_{11}/r$, $\gamma_{\cdot 1} = 100,000$ and $\gamma_{\cdot\cdot} = 10,000,000$. Each simulated Δ_{025} value was based on 100 000 Monte Carlo draws from the posterior IC distribution of interest. Table VI displays the fitted constants for different values of r (in the parameter fitting, $r = 0$ was approximated by $r = 0.001$ and $r = 1$ was approximated by $r = 0.999$ for computational stability).

r	A_r	B_r
0.0	3.09	2.22
0.1	2.93	2.27
0.2	2.78	2.26
0.3	2.62	2.25
0.4	2.45	2.15
0.5	2.25	2.12
0.6	2.03	2.05
0.7	1.79	1.93
0.8	1.61	1.89
0.9	1.13	1.15
1.0	0.073	-0.081

Table VI. Fitted parameters for the Δ_{025} function for different values of r

I.2. Moderating prior for the third order IC

The hyper parameters of the moderating prior for third order IC values are:

$$\begin{aligned}
\alpha_{111} &= \frac{q_{11\cdot}q_{1\cdot 1}q_{\cdot 11}}{q_{1\cdot\cdot}q_{\cdot 1\cdot}q_{\cdot\cdot 1}} \cdot \alpha_{\ldots} & \alpha_{011} &= \frac{q_{01\cdot}q_{0\cdot 1}q_{\cdot 11}}{q_{0\cdot\cdot}q_{\cdot 1\cdot}q_{\cdot\cdot 1}} \cdot \alpha_{\ldots} \\
\alpha_{110} &= \frac{q_{11\cdot}q_{1\cdot 0}q_{\cdot 10}}{q_{1\cdot\cdot}q_{\cdot 1\cdot}q_{\cdot\cdot 0}} \cdot \alpha_{\ldots} & \alpha_{010} &= \frac{q_{01\cdot}q_{0\cdot 0}q_{\cdot 10}}{q_{0\cdot\cdot}q_{\cdot 1\cdot}q_{\cdot\cdot 0}} \cdot \alpha_{\ldots} \\
\alpha_{101} &= \frac{q_{10\cdot}q_{1\cdot 1}q_{\cdot 01}}{q_{1\cdot\cdot}q_{\cdot 0\cdot}q_{\cdot\cdot 1}} \cdot \alpha_{\ldots} & \alpha_{001} &= \frac{q_{00\cdot}q_{0\cdot 1}q_{\cdot 01}}{q_{0\cdot\cdot}q_{\cdot 0\cdot}q_{\cdot\cdot 1}} \cdot \alpha_{\ldots} \\
\alpha_{100} &= \frac{q_{10\cdot}q_{1\cdot 0}q_{\cdot 00}}{q_{1\cdot\cdot}q_{\cdot 0\cdot}q_{\cdot\cdot 0}} \cdot \alpha_{\ldots} & \alpha_{000} &= \frac{q_{00\cdot}q_{0\cdot 0}q_{\cdot 00}}{q_{0\cdot\cdot}q_{\cdot 0\cdot}q_{\cdot\cdot 0}} \cdot \alpha_{\ldots}
\end{aligned} \tag{21}$$

where:

$$\alpha_{\ldots} = 0.5 \cdot \frac{q_{1\cdot\cdot}q_{\cdot 1\cdot}q_{\cdot\cdot 1}}{q_{11\cdot}q_{1\cdot 1}q_{\cdot 11}} \tag{22}$$

and:

$$
\begin{aligned}
q_{1\cdot\cdot} &= \frac{n_{1\cdot\cdot} + 1/2}{n_{\cdot\cdot} + 1} & q_{11\cdot} &= \frac{n_{11\cdot} + 1/4}{n_{\cdot\cdot} + 1} & q_{0\cdot1} &= \frac{n_{0\cdot1} + 1/4}{n_{\cdot\cdot} + 1} \\
q_{0\cdot\cdot} &= \frac{n_{0\cdot\cdot} + 1/2}{n_{\cdot\cdot} + 1} & q_{10\cdot} &= \frac{n_{10\cdot} + 1/4}{n_{\cdot\cdot} + 1} & q_{0\cdot0} &= \frac{n_{0\cdot0} + 1/4}{n_{\cdot\cdot} + 1} \\
q_{\cdot1\cdot} &= \frac{n_{\cdot1\cdot} + 1/2}{n_{\cdot\cdot} + 1} & q_{01\cdot} &= \frac{n_{01\cdot} + 1/4}{n_{\cdot\cdot} + 1} & q_{\cdot11} &= \frac{n_{\cdot11} + 1/4}{n_{\cdot\cdot} + 1} \\
q_{\cdot0\cdot} &= \frac{n_{\cdot0\cdot} + 1/2}{n_{\cdot\cdot} + 1} & q_{00\cdot} &= \frac{n_{00\cdot} + 1/4}{n_{\cdot\cdot} + 1} & q_{\cdot10} &= \frac{n_{\cdot10} + 1/4}{n_{\cdot\cdot} + 1} \\
q_{\cdot\cdot1} &= \frac{n_{\cdot\cdot1} + 1/2}{n_{\cdot\cdot} + 1} & q_{1\cdot1} &= \frac{n_{1\cdot1} + 1/4}{n_{\cdot\cdot} + 1} & q_{\cdot01} &= \frac{n_{\cdot01} + 1/4}{n_{\cdot\cdot} + 1} \\
q_{\cdot\cdot0} &= \frac{n_{\cdot\cdot0} + 1/2}{n_{\cdot\cdot} + 1} & q_{1\cdot0} &= \frac{n_{1\cdot0} + 1/4}{n_{\cdot\cdot} + 1} & q_{\cdot00} &= \frac{n_{\cdot00} + 1/4}{n_{\cdot\cdot} + 1}
\end{aligned}
\tag{23}
$$

III


Machine Learning, 58, 79–94, 2005

Case Based Imprecision Estimates for Bayes Classifiers with the Bayesian Bootstrap

G. NIKLAS NORÉN niklas.noren@who-umc.org;noren@math.su.se
WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden; Mathematical Statistics, Stockholm University, Stockholm, Sweden

ROLAND ORRE roland.orre@neurologic.se;orre@math.su.se
NeuroLogic, Johan Enbergs v. 28, 171 61 Solna, Sweden; Mathematical Statistics, Stockholm University, Stockholm, Sweden

Editor: Dale Schuurmans

Abstract. This article outlines a Bayesian bootstrap method for case based imprecision estimates in Bayes classification. We argue that this approach is an important complement to methods such as k-fold cross validation that are based on overall error rates. It is shown how case based imprecision estimates may be used to improve Bayes classifiers under asymmetrical loss functions. In addition, other approaches to making use of case based imprecision estimates are discussed and illustrated on two real world data sets. Contrary to the common assumption, Bayesian bootstrap simulations indicate that the uncertainty associated with the output of a Bayes classifier is often far from normally distributed.

Keywords: case based imprecision estimates, Bayes classifier, Bayesian bootstrap, naive Bayes

1. Introduction

In supervised learning, a set of labelled training examples, with known values for both predictor and response variables, is provided. The general aim is to train a classifier to predict unobserved variable values of new instances, based on the characteristics of the labelled instances.

Bayes classifiers put supervised learning in a probabilistic framework where all possible values of a missing response variable are assigned estimated probabilities, based on the values of the observed variables and on prior probabilities for the unobserved variables. This is especially useful when the response variable is not fully determined by the predictor variables, i.e. when two cases with identical values for the predictor variables may have different values for the response variable. Under such circumstances there is clearly uncertainty associated with any output of the classifier.

The accuracy of Bayes classifiers has previously been studied with methods such as k-fold cross-validation (Kohavi, 1995), which give overall accuracy estimates for a classifier given some training data. Such methods do not account for the variability between cases in the associated uncertainty, and relate to the number correctly classified cases rather than to the precision of the attributed probabilities.

For neural networks, MacKay (1992) has suggested case specific uncertainty estimates based on Bayesian inference and the assumption that the uncertainty associated with the weights in the network can be described by normal distributions (see also Bishop, 1995). This approach allows each classification performed by such a neural network to be accompanied by precision estimates.

In this article, we propose a Bayesian bootstrap method for case based imprecision estimates similar to those of MacKay, but for Bayes classifiers. For clarity of presentation, we focus on the naive Bayes classifier (Kononenko, 1990), but the methods apply equally well to generalized Bayes classifiers such as the semi-naive classifiers discussed in for example Kononenko (1991) and Domingos and Pazzani (1997).

The aims of this article is to emphasize the importance of case based imprecision estimates, to show how the Bayesian bootstrap may be used to generate precise case based estimates and to indicate how this information can be used for better informed Bayes classification.

In earlier work, Orre et al. (2000) proposed case specific precision estimates for a semi-naive Bayes classifier based on a normal approximation, and Orre and Lansner (1996) described a method for how similar uncertainty estimates may be obtained for real-valued variables.

2. The Bayesian bootstrap

Bootstrap methods in general (Efron, 1979) study how parameter estimates vary when a data set is resampled. A special type of bootstrap method is the Bayesian bootstrap (Rubin, 1981). In the Bayesian bootstrap, replicates of a given data set $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ are generated by assigning Dirichlet distributed random weights to the cases $\mathbf{z}_i$ in the original data set. The parameter of interest is calculated for each bootstrap replicate, and as shown by Rubin (1981), the distribution of the calculated parameter values over the replicated data sets approximates the posterior distribution of this parameter. This is very helpful in situations where no closed form expression for the posterior distribution is known. Based on the Bayesian bootstrap replicates of a data set, we may form a histogram for the full posterior distribution or calculate point estimates such as the posterior mean estimate or estimates for different percentiles of the distribution.

The most straightforward approach for resampling in the Bayesian bootstrap, is to assign $Di(1, \ldots, 1)_n$ distributed weights to the observed instances $\mathbf{z}_i$. However, when many $\mathbf{z}_i$ are equal, it is more efficient to assign $Di(n_1, \ldots, n_m)$ weights to the m distinct values $\mathbf{d}_j$ of $\mathbf{Z}$, where n_j is the number of $\mathbf{z}_i$ equal to $\mathbf{d}_j$. Due to the nature of the Dirichlet distribution, these two operations are mathematically equivalent.

Let $\theta = \{\theta_1, \ldots, \theta_m\}$ be the vector of probabilities $\theta_j = P(\mathbf{Z} = \mathbf{d}_j)$. With $Di(n_1, \ldots, n_m)$ distributed weights, the Bayesian bootstrap posterior distribution is proportional to:

$$\prod_{j=1}^{m} \theta_j^{n_j - 1} \tag{1}$$

and the corresponding implicit prior distribution is proportional (Rubin, 1981):

$$\prod_{j=1}^{m} \theta_j^{-1} \tag{2}$$

This is sometimes referred to as Haldane's prior.

Bayesian bootstrap simulation based on a more general prior distribution is however possible. A prior distribution proportional to:

$$\prod_{j=1}^{m} \theta_j^{l_j - 1} \tag{3}$$

yields a posterior distribution proportional (Rubin, 1981):

$$\prod_{j=1}^{m} \theta_j^{n_j + l_j - 1} \tag{4}$$

It can be simulated by assigning $Di(n_1 + l_1, \ldots, n_m + l_m)$ distributed weights to the vector of distinct values $\mathbf{d} = \{\mathbf{d}_1, \ldots, \mathbf{d}_m\}$.

A major advantage of such a more general prior distribution is that zero counts (when two variable values have never been observed together) may be handled in a better way than with Haldane's prior or with classical statistics. In fact, Domingos and Pazzani (1997) uses this argument to motivate the use of a Laplace corrector with $f = 1/n$, which is technically equivalent to a Dirichlet prior distribution with $l_j = 1/n$.

One limitation of bootstrap methods in general is the indirect assumption that all possible variable values have been observed. In fact, with several variables, the bootstrap methods effectively assume that all possible *combinations* of variable values have been observed, but in Section 4.1 we propose a modification to the Bayesian bootstrap that reduces the negative impact of this assumption.

The choice of prior distribution clearly has an impact on any analysis based on Bayesian bootstrap simulation. In the experiments presented in this article, we have aimed to minimize the prior's impact on the final result, while retaining a moderating effect, by using a prior distribution with small prior sample size (Gelman et al., 1995). The sensitivity of the results to the choice of prior has also been investigated.

3. Bayes classifiers

The aim of Bayes classifiers is to assign the value:

$$\underset{y_j}{\text{argmax}}\ P(Y = y_j \mid X_1 = x_1, \ldots, X_m = x_m) \tag{5}$$

to the response variable Y, for any unlabelled instance with predictor variable values

$$(x_1, \ldots, x_m)$$

The actual class probabilities are however unknown parameters of an underlying model, wherefore classifiers are generally based on estimates $\hat{P}(y_j \mid \mathbf{x})$ from a batch of labelled training data:

$$(\mathbf{x}(k), y(k)), \quad k = 1, \ldots, n$$

with observed values for both predictor and response variables.

Implemented Bayes classifiers often use classical maximum likelihood estimates, but the method proposed in this article is based on a full Bayesian approach. Bayesian inference combines prior information on a parameter's value with observed data, to yield the posterior probability distribution of the parameter (Gelman et al., 1995). In the following, we either consider the full posterior distribution of a parameter or use Bayesian point estimates such as the *posterior mean estimate* or the *maximum à posteriori estimate*.

If the number of training examples is large compared to the number of possible predictor variable value configurations, the outcome probabilities

$$P(y \mid \mathbf{x}) \tag{6}$$

may be directly estimated from data. Classifiers based on full conditional probabilities are sometimes referred to as optimal Bayes classifiers (Mitchell, 1997), since they rely on no assumptions of mutual independence between the different predictor variables.

The drawback for the optimal Bayes approach is that in real applications, there are seldom large enough numbers of training examples to sufficiently populate the entire domain of possible predictor variable configurations. This sparsity of data increases rapidly with the number of predictor variables used, due to the exponential increase in the number of possible configurations, something that is commonly referred to as the curse of dimensionality.

3.1. The naive Bayes classifier

In naive Bayes classification, the predictor variable values are assumed to be mutually independent conditional on the class, and this assumption allows for the following modified expression:

$$\begin{aligned} P(y \mid \mathbf{x}) &= \frac{P(\mathbf{x} \mid y)}{P(\mathbf{x})} \cdot P(y) \propto P(\mathbf{x} \mid y) \cdot P(y) \\ &= P(x_1 \mid y) \cdot \cdots \cdot P(x_m \mid y) \cdot P(y) \end{aligned} \tag{7}$$

which is normalized through division by the sum of the different class probabilities:

$$\sum_j P(y_j \mid \mathbf{x})$$

With respect to the curse of dimensionality, the main advantage of the naive Bayes approach is that it is based on estimates of marginal probabilities, $P(x_i \mid y)$, rather than of

full conditional probabilities, $P(y \mid \mathbf{x})$, something that significantly reduces the amount of training data required for reliable parameter estimation.

The drawback for the naive Bayes approach is that the underlying assumption of mutual independence between all predictor variables is commonly violated. Nevertheless, this approach has proven to be very versatile and to often compare well with more sophisticated methods (Domingos and Pazzani, 1997; Hand and Yu, 2001).

3.2. The semi-naive Bayes classifier

Semi-naive Bayes classifiers (Kononenko, 1991) allow for data models where some but not all dependencies between variables are accounted for. Groups of dependent predictor variables are encoded as composite variables whose possible values are combinations of the original variables' values. The aim is to identify a set of mutual independence assumptions that optimizes the trade-off between accuracy and computational efficiency.

Consider for example a Bayes classifier where on one hand x_1, x_2 and x_3 and on the other hand x_4 and x_5 are coencoded as composite variables. Equation (7) becomes:

$$P(y \mid \mathbf{x}) \propto P(x_1, x_2, x_3 \mid y) \cdot P(x_4, x_5 \mid y) \cdot \cdots \cdot P(x_m \mid y) \cdot P(y) \tag{8}$$

The semi-naive Bayes classifier may be regarded as a naive Bayes classifier with respect to the coencoded variables. For clarity of presentation we will therefore focus the remainder of our discussion on the naive Bayes classifier, but it should be kept in mind that all presented methods apply to other Bayes classifiers as well.

4. Methodology

Because Bayes classifiers are based on re-expressions of $P(y \mid x_1, \ldots, x_m)$ as products of several unknown probability parameters (see Eqs. (7) and (8)), no analytical form for the posterior distributions of the class probabilities is known. We propose the Bayesian bootstrap method be used to obtain accurate estimates for these posterior distributions. Given a Bayes classifier and a large enough number of bootstrap replicates, the Bayesian bootstrap yields arbitrarily accurate estimates, and unlike methods proposed in earlier work (MacKay, 1992; Orre et al., 2000), it does not rely on normal approximations.

4.1. An adjusted Bayesian bootstrap

The original Bayesian bootstrap method requires the assignment of a random weight to each individual case in the training data. For large data sets this may be intractable—to draw 10 000 bootstrap replicates from a data set with a million cases requires around 10 billion non-uniform random numbers to be generated. Clearly, under such circumstances a more efficient approach is necessary.

In principle, the computational complexity of the Bayesian bootstrap may be reduced by assigning random weights to each distinct set of predictor variable values rather than to

Table 1. Pseudo code for the adjusted Bayesian bootstrap algorithm.

- Let n_{y_j} be the number of cases in training data that are labelled y_j.
- Let $n_{x_i y_j}$ be the number of cases in training data with $X_i = x_i$ and $Y = y_j$
- Let l_{y_j} be the prior hyper parameter for $Y = y_j$
- Let $l_{x_i y_j}$ be the prior hyper parameter for $X_i = x_i$ given $Y = y_j$
- Let $l^*_{x_i y_j}$ be the prior hyper parameter for $X_i \neq x_i$ given $Y = y_j$
- Let betarnd and dirrnd denote generic random number generators for the beta and the dirichlet distributions respectively (accepting as input the hyper parameters)

For each bootstrap replicate:

% Draw bootstrap marginal probabilities
$\left[P^*(y_1), \ldots, P^*(y_k)\right] = \text{dirrnd}(n_{y_1} + l_{y_1}, n_{y_2} + l_{y_2}, \ldots)$

% Draw bootstrap conditional probabilities

For each (x_i, y_j) pair:

$P^*(x_i \mid y_j) = \text{betarnd}(n_{x_i y_j} + l_{x_i y_j}, n_{y_j} - n_{x_i y_j} + l^*_{x_i y_j})$

% Calculate unnormalized bootstrap output probabilities

For each response variable value y_j:

$P^*(y_j \mid \mathbf{x}) = P^*(y_j) \cdot \prod_i P^*(x_i \mid y_j)$

% Normalize the output probabilities

For each response variable value y_j:

$P^*_n(y_j \mid \mathbf{x}) = P^*(y_j \mid \mathbf{x}) / \sum_j P^*(y_j \mid \mathbf{x})$

Return the sets of normalized bootstrap output probabilities (one set for each replicate)

each specific case in the training data set, as discussed in Section 2. However, the number of predictor variables is in practice often large enough that there is only a small number of cases with the exact same sets of predictor variable values. Consequently, this may not be sufficient to make the Bayesian bootstrap computationally tractable.

One way to further decrease the computational complexity of the Bayesian bootstrap is to incorporate the mutual independence assumptions on which the Bayes classifier relies into the resampling procedure. In such an adjusted Bayesian bootstrap approach, each factor in the Bayes classifier formula is simulated independently, and bootstrap replicates are generated as indicated in Table 1.

The adjusted Bayesian bootstrap method produces the posterior class distribution under the given model assumptions (i.e., it accounts for the mutual independence assumptions in the resampling). This further reduces the impact of the bootstrap assumption that all possible combinations of variable values have been observed (see Section 2). By simulating all predictor variables separately, any two variable values that occur separately in the training data are assumed to have a positive probability to cooccur.

Furthermore, this adjustment to the Bayesian bootstrap facilitates the assertion of a prior distribution. In the adjusted Bayesian bootstrap, each variable has its own prior distribution, so the number of prior parameters is equal to $\sum_i v_i$ (where v_i is the number of distinct variable values for variable X_i) instead of to $\prod_i v_i$ as in the original Bayesian bootstrap.

4.2. How to make use of the posterior class probability distributions

There are several ways to put the Bayesian bootstrap distributions to use. Detailed information about imprecision in the output probabilities of a Bayes classifier may be used to test model assumptions such as e.g. normal approximations, and to investigate what factors influence the imprecision in Bayes classification.

MacKay (1992) proposes marginalization (averaging) over the posterior distribution of a classification as a way to moderate the output of a neural network. For binary variables, this tends to pull the output probabilities toward 0.5, and the effect is stronger the less training data there is available. In effect this corresponds to the use of posterior means instead of maximum likelihood estimates, and it would be straightforward to implement the same principle for Bayesian bootstrap analysis of Bayes classifiers. A slightly generalized approach is to assert a baseline probability for each class, and to only output a different value (the closest credibility interval limit) for the probability if the credibility interval excludes the baseline value. This allows for a stronger moderating effect, which can be fine tuned by varying the coverage of the credibility interval. It also allows for moderation toward other values than 0.5 between 0 and 1. Another approach is to altogether refrain from making classifications with too little support in data, and instead flag cases as uncertain if the posterior interval spans the prior probabilities.

A situation where detailed information about the output class probability posterior distributions may be particularly useful is in Bayes classification under asymmetrical loss functions. Bayes classifiers typically output the most probable class for an unlabelled case, but if the different types of misclassifications have different associated losses, this is suboptimal. For binary classifiers and loss functions based on variable misclassifications costs, classification based on percentiles equal to the ratios of the misclassification costs minimizes the expected loss. Section 5.3 presents a detailed example of this. Another example is filters for unwanted e-mail messages (spam), where it is generally more severe to misclassify a wanted e-mail message as *spam* than to misclassify spam as *wanted*.

5. Examples

To illustrate the usefulness of the proposed approach for real data, we have applied it to data sets from the UCI machine learning repository (Blake & Merz, 1998). The results are presented in this section.

5.1. Setup of experiments

Two data sets from the UCI machine learning repository were selected: the mushrooms data set and the zoology data set. From the mushrooms data set, we excluded the predictor variable *veil-type* for which only one value (*partial*) is ever observed. Some of the analyses were based on subsets of the available cases and/or predictor variables in the data sets, in order to better illustrate the impact of uncertainty on the classification.

For the prior distribution of predictor variable X_i conditional on the response variable value y_j (see Table 1), we used hyper parameters $l_{x_i y_j} = \frac{1}{v_i}$ and $l^*_{x_i y_j} = \frac{v_i - 1}{v_i}$ where v_i is the

number of distinct values for X_i as before. For the response variable this amounts to hyper parameters $l_{y_j} = 1$. This is a low impact prior distribution with prior sample size v_r equal to the number of possible values for the response variable Y.

5.2. *Precise posterior distribution estimation*

To illustrate how the Bayesian bootstrap may be used to infer detailed knowledge about the posterior distribution of the Bayes classifier's output, we have used precise Bayesian bootstrap simulations (1000 replicates) to study the posterior distributions of different naive Bayes classifiers applied to the mushrooms data set.

All distributions in figure 1 are produced by the same naive Bayes classifier, which uses the following four predictor variables: *cap shape*, *cap surface*, *cap color* and *bruises*. The

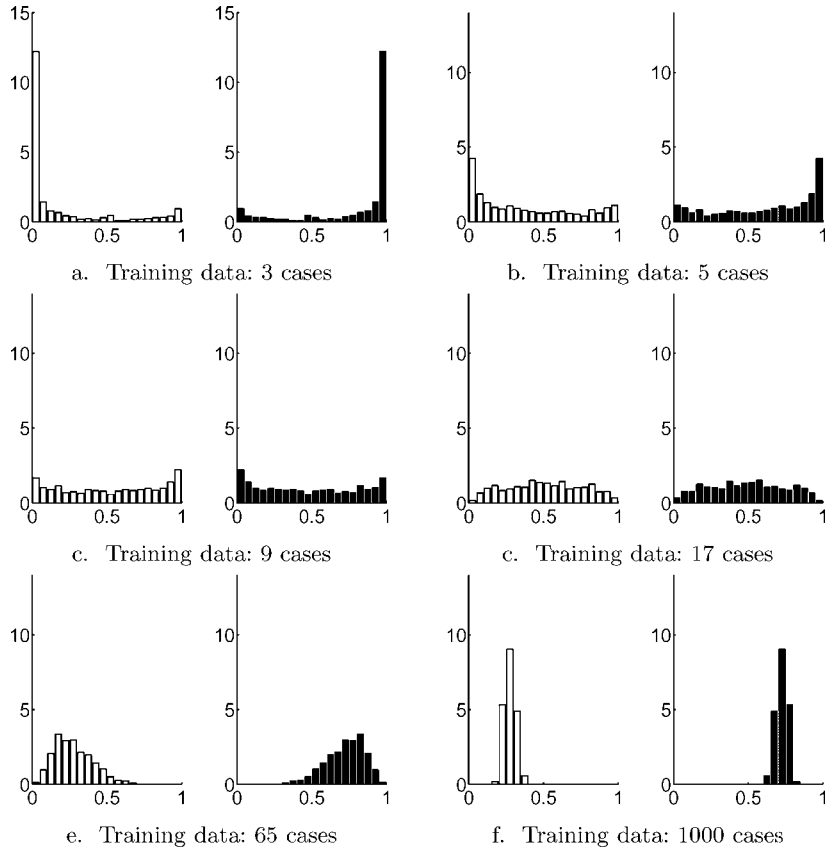


Figure 1. Posterior distributions for the probability that mushroom number 8124 in the UCI ML repository data set is edible (the leftmost distribution in each pair) and poisonous (rightmost) for varying amounts of training data, using the first 4 attributes (*cap shape*, *cap surface*, *cap color* and *bruises*).

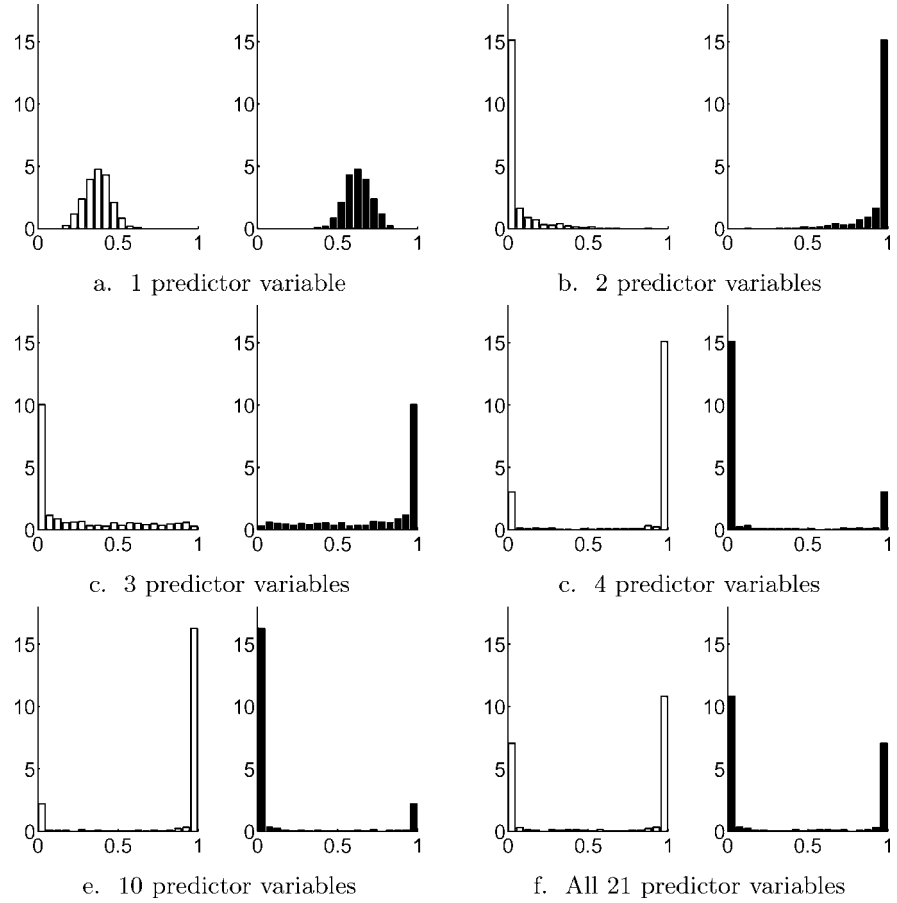


Figure 2. Posterior distributions for the probability that mushroom number 8124 in the UCI ML repository data set is edible (the leftmost distribution in each pair) or poisonous (rightmost) for varying numbers of predictor variables, based on 80 training cases. The predictor variables were added in the following order: *stalk shape*, *population*, *odor*, *stalk color below ring*, *cap color*, *gill spacing*, *stalk surface below ring*, *ring number*, *ring type* and then the rest.

aim is to show how the uncertainty in the output decreases as more training data is added. Please note how the shape of the posterior distribution transforms from its bathtub shape for small amounts of training data over an almost uniform distribution for intermediate amounts of training data to a more normal-like posterior distribution for large amounts of training data.

In figure 2 we have used constant training data, but different naive Bayes classifiers. The difference between the classifiers is the number of predictor variables on which they are based, and the aim is to illustrate how the uncertainty in the output increases as more predictor variables are added.

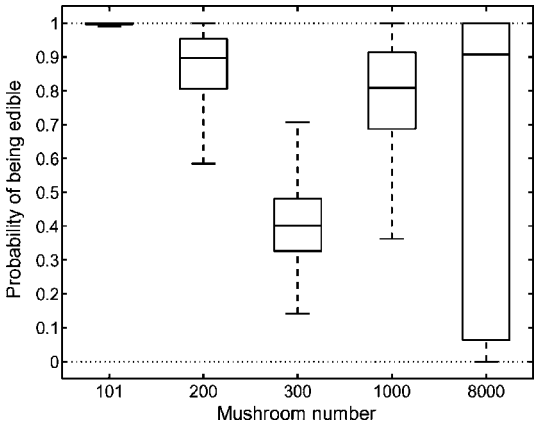


Figure 3. Box plots for the output classification of different mushrooms based on the predictor variables: *cap shape*, *cap surface*, *cap color* and *bruises* and the first 100 mushrooms in the data set. Clearly, the degree of uncertainty varies for the five mushrooms, even though the same Bayes classifier has been used.

5.3. Classifying mushrooms with variable misclassification costs

As an illustration of how case based precision estimates may be incorporated into a Bayes classifier, a naive Bayes classifier was trained on the first 100 cases in the mushrooms data set using the first four predictor variables: *cap shape*, *cap surface*, *cap color* and *bruises*. Figure 3 displays box plots for the uncertainty associated with the classifications of five different mushrooms that were not included in the training data.

If the loss associated with the misclassification of a poisonous mushroom as edible is higher than the loss associated with the misclassification of an edible mushroom as poisonous, simple naive Bayes classification (outputting the class estimated to be the most likely) is suboptimal. Assume for simplicity that the loss associated with the former type of misclassification is twice that of the latter; without case based imprecision estimates, the easiest way to account for this asymmetry is to change the cut-off from 50 to $66.\bar{6}\%$ (i.e., only label a mushroom as edible if the point estimate for this probability is greater than 2/3, since this is the level at which the expected loss of classifying the mushroom as *edible* is the same as that of classifying the mushroom as *poisonous*). The reliability of the naive Bayes probability estimates is however questionable as the naive Bayes classifier tends to over-estimate the confidence in its predictions (Hand and Yu, 2001). With case based uncertainty estimates, an alternative approach is to instead consider variation in the output classification (with cut-off 50%) over the Bayesian bootstrap distribution and only label the mushroom as *edible* if more than 2/3 of the bootstrap replicates indicate *edible*.

To compare these two approaches, we have used five different naive Bayes classifiers each trained on 100 out of the first 500 mushrooms in the data set (and the predictor variables: *cap shape*, *cap surface*, *cap color* and *bruises*) to classify the last 100 mushrooms in the

Table 2. The efficiency of three naive Bayes decision rules, for five different naive Bayes classifiers trained on subsets of the UCI mushrooms data set.

Training data	True pos.	False pos.	Sens.	Spec.	Loss
		a. $\hat{P}(edible) > 1/2$			
1–100	54	33	1.00	0.28	66
101–200	54	25	1.00	0.46	50
201–300	54	25	1.00	0.46	50
301–400	54	13	1.00	0.72	26
401–500	40	4	0.74	0.91	22
Averages	51.2	20.0	0.95	0.57	42.8
		b. $\hat{P}(edible) > 2/3$			
1–100	54	25	1.00	0.46	50
101–200	54	12	1.00	0.74	24
201–300	40	5	0.74	0.89	24
301–400	40	4	0.74	0.91	22
401–500	40	3	0.74	0.93	20
Averages	45.6	9.8	0.84	0.79	28.0
		c. $P(P^*(edible) > 1/2) > 2/3$			
1–100	43	4	0.80	0.91	19
101–200	39	2	0.72	0.96	19
201–300	38	2	0.70	0.96	20
301–400	40	3	0.74	0.93	20
401–500	40	2	0.74	0.96	18
Averages	40.0	2.6	0.74	0.94	19.2

data set. A comparison of the two decision rules and the standard naive Bayes decision rule is displayed in Table 2.

5.4. Sensitivity to the choice of prior

To evaluate how sensitive the Bayesian bootstrap method is to variations in the choice of prior distribution, we have compared the uncertainty estimates for mushroom 8124 with the chosen prior ($l_{x_i y_j} = \frac{1}{v_i}$, $l^*_{x_i y_j} = \frac{v_i - 1}{v_i}$ and $l_{y_j} = 1$) to uncertainty estimates based on two other data sensitive priors: the uniform prior ($l_{x_i y_j} = l^*_{x_i y_j} = l_{y_j} = 1$) and Haldane's prior ($l_{x_i y_j} = l^*_{x_i y_j} = l_{y_j} = 0$). The results are displayed in figure 4.

5.5. Function approximation

Figures 5 and 6 display bootstrap posterior distributions for cases in the UCI zoology and mushrooms data sets, together with fitted beta distributions.

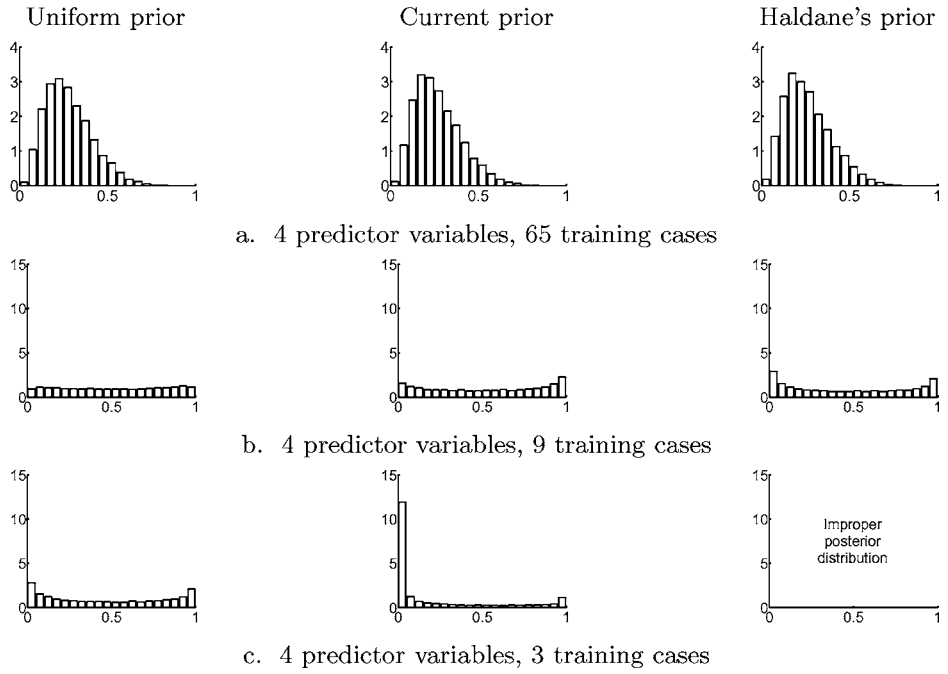


Figure 4. Sensitivity to the choice of prior hyper parameters of the uncertainty estimates for the classification of mushroom 8124 based on various amounts of training data.

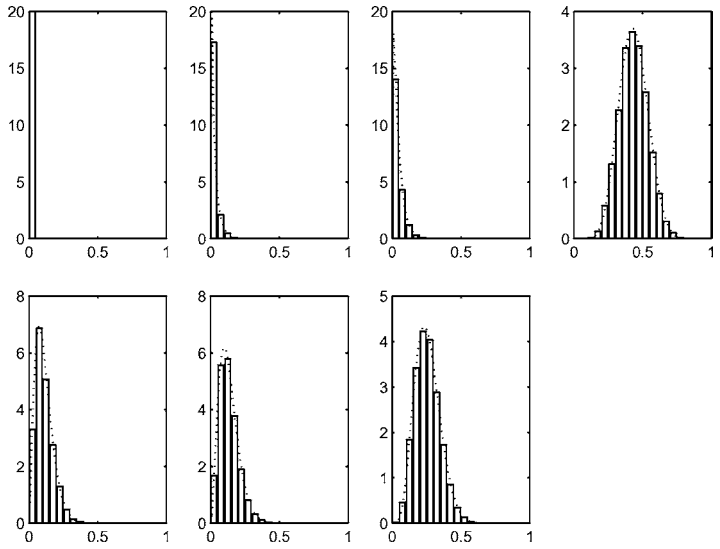


Figure 5. Fitted beta distributions (dotted curves) for the 7 output distributions (one for each class) of animal number 81 in the UCI zoology data set, based on the first 80 animals in the data set and the first four predictor variables.

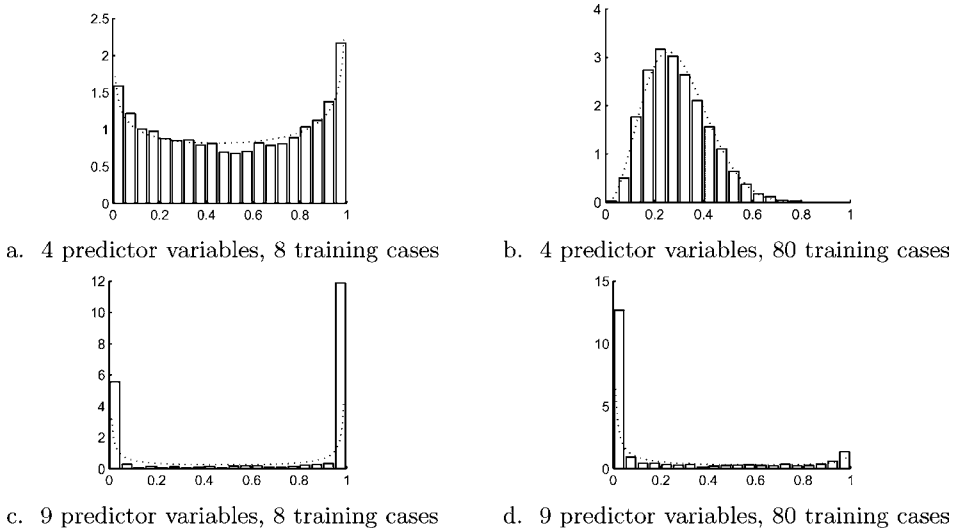


Figure 6. Fitted beta distributions (dotted curves) for the probability of mushroom number 8124 to be edible, with different numbers of predictor variables and varying amounts of training data.

6. Discussion

The results reported on in Section 5.3 indicate that case based precision estimates may allow for more robust decision rules when the loss function is asymmetrical. Table 2 shows that over a rather large domain in the UCI mushrooms data set (five classifiers trained on different portions of the database and each applied to 100 cases) the Bayesian bootstrap based decision rule has better specificity, and lower average loss for variable misclassification costs. A plausible explanation for this is the tendency of the naive Bayes classifier to "probability overshoot", i.e. to overestimate the confidence in its predictions (Hand and Yu, 2001). Figure 2 illustrates this: even when there is significant uncertainty due to a large number of predictor variables compared to the amount of training data available, the most likely value of the output probability tends to 0 or 1. Due to this low reliability of the naive Bayes probability estimates, even crude Bayesian bootstrap simulation may yield better uncertainty estimates for Bayes classifiers.

However, case based imprecision estimates focus solely on imprecision due to limits in the amount of relevant training data available, and do not account for erroneous assumptions in the design of the classifier. Other methods such as k-fold cross-validation should therefore always be used to test the accuracy of a Bayes classifier. If the accuracy of the classifier is poor, the Bayesian bootstrap may be used to deduce whether this could be due to limits in the amount of relevant training data, or whether it is solely attributable to incorrect model assumptions.

A good example of this distinction between precision and accuracy is the classification of mushroom number 8124 in figure 1. As more and more mushrooms are added

to the training data, the imprecision in the classification is gradually reduced, and with 1000 mushrooms in the training data, the output probabilities are quite precise (centered at around $P(poisonous) = 0.7$). However, if all predictor variables are used, the output for $P(poisonous)$ is instead very close to 0, and indeed the true label of mushroom number 8124 is *edible*. Clearly the problem here is not the amount of relevant training data, but the design (e.g. the feature selection) of the Bayes classifier.

The experiments reported on in this article are in agreement with the observation by Hand and Yu (2001) that the more predictor variables that are used in the Bayes classifier, the more training data is required for reliable prediction. This relates to the issue of irrelevant variables: if a predictor variable is completely unassociated with the response variable, its average effect will be to only increase the uncertainty in every prediction. This is a good incentive for identifying and excluding from the analysis any irrelevant variables.

Instead of the Bayesian bootstrap, the original bootstrap (Efron, 1979) could be used to generate case based precision estimates (based on sampling distributions for parameter estimates rather than on posterior distributions for the parameters). However, the non-Bayesian bootstrap does not allow for the use of prior distributions, which as discussed in Section 4.1 may reduce the negative impact of the bootstrap assumption that all possible data points have been observed. In addition, the parameter estimate distribution is discrete, and may be inconsistent with the observed data: consider for example an attempt to study the probability p of a binomial distribution with the original bootstrap. With four observations—two successes and two failures—each bootstrap replicate has a 1/16 chance of yielding $\hat{p}^* = 1$ for the probability of success, which is clearly in disagreement with data since under this model, the probability of observing the two failures is 0.

As discussed in Section 4.1, the proposed adjustment to the Bayesian bootstrap helps to further reduce the negative impact of the bootstrap assumption that all possible data points have been observed. In addition, it often allows for more efficient simulation: if v_i is the number of distinct variable values for variable X_i, then the computational complexity of the original Bayesian bootstrap is proportional to $\alpha \cdot \prod_i v_i$, whereas the computational complexity of the adjusted Bayesian bootstrap is proportional to $\sum_i v_i$ (α is a factor that relates to how many of the possible variable value combinations that actually occur in the data set).

It is difficult to give general guidelines for how many Bayesian bootstrap replicates are required for a given reliability. To a large extent, this depends on the specific purpose for which the Bayesian bootstrap is carried out—the expected number of replicates for accurate simulation of the posterior mean is for example lower than for the 0.01 quantile (Gelman et al., 1995). For standard purpose simulations a pragmatic approach may be to, in advance, study the sampling variability of the Bayesian bootstrap estimates for a given statistic and a given number of replicates.

The main drawback of the Bayesian bootstrap approach is the computational complexity. Analytical expressions based on normal approximations have been proposed, but the results in Section 5 indicate that the true output of a Bayes classifier is often far from normally distributed. On the other hand, the results in Section 5.5 suggest that fitted beta distributions often provide good approximations, and it would be a great advantage if approximate closed form expressions for the hyper parameters of the best fit beta distribution could be derived.

Computationally intense Bayesian bootstrap simulations could then be replaced by simple formulae for the hyper parameters of a beta distribution. Clearly, this is a highly relevant area for future research related to this article.

The Bayesian bootstrap is by definition based on Dirichlet priors, but Monte Carlo simulation based on a different data model could be carried out and may generate different results. The choice of hyper parameters clearly affects the final output distribution, but as figure 4 indicates, the output class probabilities of a Bayes classifier are rather insensitive to such changes, as long as fairly weak priors are used and as long as training data consists of more than just a few cases. The prior used throughout this article typically has the bath tub shape (its exact shape depends on the number of variables and the numbers of variable values), but it may be argued that a better approach is to set the prior parameters so that the prior distribution for the class probability is always uniform. This would however result in a larger prior sample size, and a less data sensitive Bayes classifier.

7. Conclusions

The usefulness of case based uncertainty estimates for Bayes classifiers was demonstrated on real world data. It was shown how detailed information about the posterior distributions of the class probabilities may allow for better informed decisions and improve classification under asymmetrical loss functions. Contrary to assumptions of previous models, Bayesian bootstrap simulations indicate that the posterior class probability distributions of Bayes classifiers are often far from normally distributed.

Acknowledgments

The authors would like to thank A. Bate and E. Swahn for helpful comments on earlier drafts of this article. In addition, the authors would like to thank those who contributed the data sets used in the empirical studies (please see the documentation of the UCI machine learning repository for details).

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Blake, C., & Merz, C. (1998). UCI Repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Domingos, P., & Pazzani, M. (1997). On the optimatility of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–130.

Efron, B. (1979). Bootstrap methods Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall.

Hand, D. J., & Yu, K. (2001). Idiot's Bayes—Not so stupid after all? *International Statistical Review, 69:3*, 385–398.

Kohavi, R. (1995). 'A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1137–1145).

Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga, J. Boose, B. Gaines, G. Schreiber, & M. van Someren (Eds.), *Current trends in knowledge acquisition*. IOS Press.

Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Proceedings of the Sixth European Working Session on Learning* (pp. 206–219). Springer.

MacKay, D. J. (1992). The evidence framework applied to classification networks. *Computation and Neural Systems, 4:5*, 698–714.

Mitchell, T. M. (1997). *Machine Learning* 1st edition. McGraw-Hill.

Orre, R., & Lansner, A. (1996). Pulp quality modelling using Bayesian mixture density neural networks. *Journal of Systems Engineering, 6*, 128–136.

Orre, R., Lansner, A., Bate, A., & Lindquist, M. (2000). Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics & Data Analysis, 34*, 473–493.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics, 9:1*, 130–134.

Received March 18, 2003
Revised July 6, 2004
Accepted July 20, 2004
Final manuscript August 12, 2004