Data Min Knowl Disc (2007) 14:305–328
DOI 10.1007/s10618-006-0052-8

Duplicate detection in adverse drug reaction surveillance

G. Niklas Norén · Roland Orre · Andrew Bate · I. Ralph Edwards

Received: 22 December 2005 / Accepted: 26 May 2006 / Published online: 7 February 2007

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 (ADR) incidents that occur after drugs are on the market. The presence of duplicate case reports is an important data quality problem and their detection remains a formidable challenge, especially in the WHO drug safety database where reports are anonymised before submission. In this paper, we propose a duplicate detection method based on the hit-miss model for statistical record linkage described by Copas and Hilton, which handles the limited amount of training data well and is well suited for the available data (categorical and numerical rather than free text). We propose two extensions 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, and we demonstrate the effectiveness both at identifying the most likely duplicate for a given case report (94.7% accuracy) and at discriminating true duplicates from random matches (63% recall with 71% precision). The

Responsible editor: Hannu Toivonen.

G. N. Norén (✉) · A. Bate · I. R. Edwards
WHO Collaborating Centre for International Drug Monitoring,
Uppsala, Sweden
e-mail: niklas.noren@who.umc.org

G. N. Norén
Mathematical Statistics, Stockholm University,
Stockholm, Sweden

R. Orre
NeuroLogic Sweden AB,
Stockholm, Sweden

proposed method allows for more efficient data cleaning in post-marketing drug safety data sets, and perhaps other knowledge discovery applications as well.

Keywords Data cleaning · Duplicate detection · Hit-miss model

1 Introduction

The nature of pharmaceutical development means that some adverse drug reactions (ADRs) (Edwards and Aronson 2000) will only be discovered after drug approval (Evans 2000). Systems that collect case reports of suspected ADR incidents from health care professionals are referred to as spontaneous reporting systems and are a core component of post-marketing surveillance (Edwards 1999). 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 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 79 member countries of the WHO Programme for International Drug Monitoring routinely forward ADR case reports submitted to their national pharmacovigilance centres to the Uppsala Monitoring Centre. The first case reports in the WHO database date back to 1967 and as of November 2005 there are over 3.5 million reports in total in the data set; currently around 200,000 new reports are added to the database each year.

Data cleaning is an early but essential step in the knowledge discovery process (Fayyad et al. 1996; Hernández and Stolfo 1998) with the aim of improving data quality. Good data quality is in turn a prerequisite for useful data analysis (Kim et al. 2003; De Veaux and Hand 2005). The analysis of spontaneous case reports is one of the most important methods for discovering previously unknown safety problems after drugs are on the market (Rawlins 1988), but it is sometimes impaired by poor data quality (Lindquist 2004), and in particular by the presence of duplicate case reports. Quantitative methods are important in screening spontaneous reporting data for new drug safety problems (Bate et al. 1998), and may highlight potential problems based on as few as 3 case reports on a particular event, so the presence of even a few duplicates may severely affect their efficacy. While there is a general consensus that duplicate case reports are a major problem in spontaneous reporting data, there is a lack of published research on the extent of the problem. A study on vaccine adverse events data quoted rates of duplication of around 5% (Nkanza and Walop 2004). However, at times the frequency may be much higher such as in the recent review of suspected quinine induced thrombocytopenia, where FDA researchers identified 28 of the 141 US case reports (20%) as duplicates (Brinker and Beitz 2002). Whereas previous work on knowledge discovery in ADR surveillance has focused exclusively on data analysis (Bate et al. 1998; Evans 2000; Orre et al. 2000; Norén et al. 2006), this paper focuses on data quality by proposing a method to highlight suspected case report duplication.

There are at least two common causes for duplication in post-marketing drug safety data: different sources (health professionals, national authorities, companies) providing separate case reports related to the same event and 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 solutions that allow case reports to be sent very easily between organisations (Edwards 1997). 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.

Naturally, duplicate case reports tend to be much more similar than non-duplicates, but there are important exceptions. For example, separate case reports may be provided for the same patient based on the same doctor's appointment if the patient has suffered from separate adverse events considered to be unrelated. Such case reports may match perfectly on date, age, gender, country and drug substances, but are not true 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, for which the patient information may differ widely depending on whether it relates to the mother or the child.

Methods to detect duplicate case reports in post-marketing drug safety data are clearly needed (Bortnichak et al. 2001). In the WHO database, the most informative record fields for matching are best interpreted as categorical or numerical, but the duplicate detection literature focuses primarily on free text matching (Monge and Elkan 1997; Sarawagi and Bhamidipaty 2002; Bilenko and Mooney 2003a). Statistical record linkage research provides a general framework for matching based on likelihood ratios (Newcombe and Kennedy, 1962; Fellegi and Sunter, 1969), which applies to different types of data and can be implemented in various ways (see for example Jaro 1989). A common problem with implementations of this general framework is that matches in a given record field are rewarded equally, regardless of the frequency for the matching event, which may be inappropriate since chance matches are more likely on common than on rare events. The hit-miss model proposed by Copas and Hilton (1990) is a latent variable model that does account for the event frequency without being overly sensitive to limitations in the amount of labelled training data available. However, the hit-miss model relies on an assumption of independence between observed events that may lead to false positives driven by matched correlated record fields (see Sect. 2.3).

In this paper, we propose a computationally efficient approach to compensate for correlated record fields and a hit-miss mixture model for robust matching of numerical record fields. An extended hit-miss model is implemented on the WHO drug safety database and demonstrated to be useful for real world duplicate detection. This paper is an expanded version of Norén, Orre and Bate (2005), and now includes new results from a database wide screen for duplicates in the entire WHO drug safety database. Additionally, a method for

linking together larger groups of duplicates, the phenomenon of unmatchable reports and how to implement hit-miss mixture model matching for free text record fields are discussed. The new results demonstrate the feasibility of the proposed method for large scale duplicate detection and provide insight into the extent and characteristics of suspected case report duplication in international ADR surveillance.

2 Methods

2.1 The hit-miss model

The hit-miss model is a probability model for how discrepancies occur between database records that relate to the same underlying event (Copas and Hilton 1990). Let $X = j$ and $Y = k$ denote the observed values for a certain record field on two different database records. Let p_j and p_k denote the associated probabilities. The joint probability for this pair of values under the independence assumption equals the product of p_j and p_k. The hit-miss model provides an estimate p_{jk} for the same probability under the assumption that the two records relate to the same event. The contribution from each record field to the total match score (its weight) 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, under the assumption of independence, the total match score is found by adding together the weights for the different record fields.

Under the hit-miss model, each observed record field X is based on a true but unobserved event $T = t$. Observed values on different records are assumed to have been generated in independent random processes resulting in a miss with probability a, a blank with probability b and a hit with probability $1 - a - b$ (see Fig. 1). For a miss X is independent of T but follows the same distribution, for a blank the value of X is missing and for a hit $X = t$.

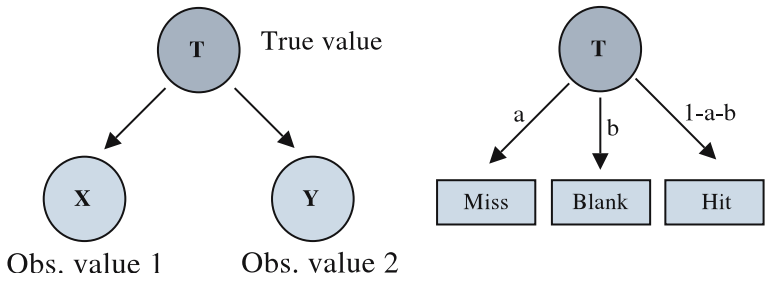

Fig. 1 The hit-miss model

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:

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

Under the hit-miss model:

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

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

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

Based on (4):

$$P(X = j) = (1 - b) \cdot \beta_j \tag{5}$$

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

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

Thus, for a given record field, we estimate b by the relative frequency of blanks in the entire database and β_i by the relative frequency of value i among non-blanks in the entire database for this record field. 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 $1 - \sum_i \hat{\beta}_i^2$.
(3), (4) and (5) give:

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

Thus, all mismatches for a given record field receive the same weight and blanks receive weight 0. It can be shown that mismatches always receive negative weights and that matches receive positive weights, as would intuitively be expected. Moreover, matches on rare events receive greater weights than matches on common events (W_{jj} decreases when β_j increases). The detailed behaviour of W_{jj} as a function of β_j is illustrated in Fig. 2 for different values of a and b.

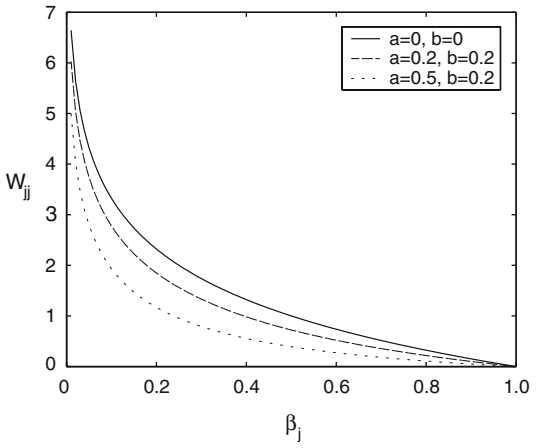

Fig. 2 $W_{jj}(\beta_j)$ based on (8), for different combinations of a and b

2.2 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 likely to yield small numerical differences between observed and true values. If, for example, two different sources provide separate case reports related to the same incident, the dates of onset may not match perfectly but are more likely to differ by a few days than by several years. Similarly, the registered patient age may differ from the true value, but small deviations are more likely than large ones. At the same time, there are other types of errors (e.g. typing errors) for which large numerical differences are as likely as small ones. In order to handle both possibilities, we propose a hit-miss mixture model which includes both 'misses' and 'deviations'. Given the true but unobserved value $T = t$, 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 $1 - a_1 - a_2 - b$ (see Table 3). 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, $X = t$ (Fig. 3).

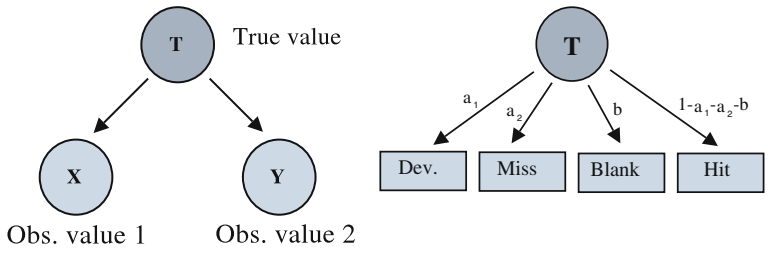

Fig. 3 The hit-miss mixture model

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

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 $f(d)$

For two observed numerical values $X=i$ and $Y=j$, we focus on the difference $d=j-i$. With respect to this, there are six different outcomes of the hit-miss mixture model as listed in Table 1 where $\phi(d;\mu,\sigma^2)$ denotes the normal probability density function with mean μ and variance σ^2 and where $\delta(d)$ 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 (e.g. a hit and a miss). Under the assumption that $\mathrm{var}(T) \gg \sigma_1^2$, the difference between a miss and a deviation approximately follows this distribution as well.

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

$$p_r(d) = \begin{cases} 1-(1-b)^2 & d \text{ missing} \\ (1-a_1-a_2-b)^2 \cdot \delta(d) + a_2(2-a_2-2b)\cdot f(d) + & d \text{ numerical} \\ +2a_1(1-a_1-a_2-b)\cdot\phi(d;0,\sigma_1^2) + a_1^2\cdot\phi(d;0,2\sigma_1^2) & \end{cases} \tag{9}$$

For unrelated records, d follows the more simple distribution:

$$p_u(d) = \begin{cases} 1-(1-b)^2 & d \text{ missing} \\ (1-b)^2\cdot f(d) & d \text{ numerical} \end{cases} \tag{10}$$

and we can calculate log-likelihood ratio based weights $W(d_1,d_2)$ by integrating (9) and (10) over an interval $[d_1,d_2]$ corresponding to the precision of d (e.g. for two observed ages 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 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.

Table 2 EM algorithm for hit-miss mixture model fitting

1. Given estimates for b and $f(d)$, make initial guesses $\hat{a}_1$, $\hat{a}_2$ and $\hat{\sigma}_1^2$

2. Calculate $\hat{\alpha}_1, \ldots, \hat{\alpha}_4$:
$$\hat{\alpha}_1 = (1 - \hat{a}_1 - \hat{a}_2 - \hat{b})^2$$
$$\hat{\alpha}_2 = \hat{a}_2(2 - 2\hat{b} - \hat{a}_2)$$
$$\hat{\alpha}_3 = 2\hat{a}_1(1 - \hat{a}_1 - \hat{a}_2 - \hat{b})$$
$$\hat{\alpha}_4 = \hat{a}_1^2$$

3. For each observed d_i in training data, compute the probability that it belongs to each mixture component
$$\hat{\gamma}_1(d_i) = \frac{\hat{\alpha}_1 \delta(d_i)}{\hat{\alpha}_1 \delta(d_i) + \hat{\alpha}_2 f(d_i) + \hat{\alpha}_3 \phi(d_i; 0, \hat{\sigma}_1^2) + \hat{\alpha}_4 \phi(d_i; 0, 2\hat{\sigma}_1^2)}$$
$$\hat{\gamma}_2(d_i) = \frac{\hat{\alpha}_2 f(d_i)}{\hat{\alpha}_1 \delta(d_i) + \hat{\alpha}_2 f(d_i) + \hat{\alpha}_3 \phi(d_i; 0, \hat{\sigma}_1^2) + \hat{\alpha}_4 \phi(d_i; 0, 2\hat{\sigma}_1^2)}$$
$$\hat{\gamma}_3(d_i) = \frac{\hat{\alpha}_3 \phi(d_i; 0, \hat{\sigma}_1^2)}{\hat{\alpha}_1 \delta(d_i) + \hat{\alpha}_2 f(d_i) + \hat{\alpha}_3 \phi(d_i; 0, \hat{\sigma}_1^2) + \hat{\alpha}_4 \phi(d_i; 0, 2\hat{\sigma}_1^2)}$$
$$\hat{\gamma}_4(d_i) = \frac{\hat{\alpha}_4 \phi(d_i; 0, 2\hat{\sigma}_1^2)}{\hat{\alpha}_1 \delta(d_i) + \hat{\alpha}_2 f(d_i) + \hat{\alpha}_3 \phi(d_i; 0, \hat{\sigma}_1^2) + \hat{\alpha}_4 \phi(d_i; 0, 2\hat{\sigma}_1^2)}$$

4. Update the variance estimate $\hat{\sigma}_1^2$:
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n \hat{\gamma}_3(d_i) \cdot d_i^2 + \hat{\gamma}_4(d_i) \cdot d_i^2/2}{\sum_{i=1}^n \hat{\gamma}_3(d_i) + \hat{\gamma}_4(d_i)}$$

   Update $\hat{a}_1$ and $\hat{a}_2$ by numerical maximisation of the total expected likelihood according to (9) over eligible parameter value pairs (such that $\hat{a}_1 + \hat{a}_2 < 1 - \hat{b}$).

5. Iterate 2–4 until convergence

2.3 A method to handle correlated record fields

The assumption of independence between record fields in the standard hit-miss model allows the total match score to be calculated by summation over the individual record field weights. The independence assumption may, however, lead to over-estimated evidence that two records are duplicates, if the matching record fields are correlated. Clearly, this may hinder effective duplicate detection in data sets where the independence assumption is not appropriate for all record fields.

To reduce the risk for high match scores driven by correlated record fields, we propose a model that accounts for pairwise associations. Let $j_1, \ldots, j_m$ denote a set of events for record fields $X_1, \ldots, X_m$. In the independence model, the probability that these events should co-occur on a database record is:

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

Under the assumption that the information in different record fields can be considered independently, the total match score contribution is:

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

When the independence assumption is not appropriate, the joint probability for the set of events $j_1, \ldots, j_m$ can be expressed as:

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

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

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

For correlated record fields, the joint distribution may be modelled by (14) instead of (11). Let:

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

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

Then:

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

and:

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

Thus, the adjusted match score can be approximated by subtracting a sum of compensating terms from the original match score, where each compensating term can be written on the following form:

$$\log_2 \frac{\beta^*_{j_t}}{\beta_{j_t}} = \log_2 \frac{P(j_t|j^*_t)}{P(j_t)} \tag{19}$$

A shrinkage estimate for this log-ratio has earlier proven useful, as a robust measure of association in screening the WHO drug safety database for interesting quantitative associations (Bate et al. 1998; Orre et al. 2000; Norén et al. 2006). It is referred to as the Information Component (IC) and is defined as:

$$\mathrm{IC}_{ij} = \log_2 \frac{P(j \mid i)}{P(j)} \tag{20}$$

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). The advantage of IC values over raw observed-to-expected ratios is that they provide less volatile estimates when little data is available. In order to provide more robust compensation for correlated record fields, we use IC shrinkage estimates for $\log_2 \frac{\beta^*_{j_t}}{\beta_{j_t}}$ in (18). Because we only ever condition on preceding events in the sequence, the ordering of events $j_1, \ldots, j_m$ affects the compensating term in (18). As a less arbitrary choice of ordering, we re-arrange the events in decreasing order of maximal IC value within the set of matched events.

3 Implementation

3.1 Data pre-processing

Although the WHO database allows for the transmission and storage of large amounts of data for each individual case report, few case reports have even the majority of the fields filled in (Bate et al. 1998). For the identification of possible duplicate records, the following record fields were considered the most informative: date of onset, patient age, patient gender, country of origin, outcome, drug substances used and ADR terms observed (drug substances and ADR terms are in fact large 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, 1184 records (0.04% of the database) have different onset dates for different ADR terms; for those records, the earliest listed onset date was consistently used. Because some countries encode dates with missing days as the first of the month, and dates with missing months and days as the first of the year, all such dates were re-encoded as partially missing (for example all occurrences of 2002-03-01 were re-encoded as 2002-03-? and 2002-01-01 as 2002-?-?). 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

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

Record field	Interpretation	Type	Missing data (%)
Date	Date of onset	Numerical (days since reference date)	23
Outcome	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

of non-standard symbols were interpreted as missing values and re-encoded as such. Different age units had been used so in order to harmonise, all ages were re-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 ignored.

For very large data sets, it may be computationally intractable to score all possible record pairs. A common strategy to reduce computational complexity is to group the records into different blocks based on their values for a subset of important record fields and to only score records within the same block (Fellegi and Sunter 1969). For the WHO database, we implicitly block on drug substances crossed with ADR categories, by only computing match scores for those record pairs that have at least one drug substance in common and share at least one ADR category (as defined by the System Organ Class, which is a higher level grouping of ADR terms). In addition to the improvement in computational efficiency, this also reduces the risk for false leads by non-duplicate case reports on different reactions in the same patient (see Sect. 1). Blocking may in theory yield extra false negatives, but duplicate records that don't match on at least one drug substance and an ADR type are unlikely to receive high enough match scores to exceed the threshold for manual review.

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

The majority of the hit-miss model parameters are estimated based on the entire data set (here, the contents of the WHO database as of June 2003), but c for categorical record fields and a_1 and a_2 for numerical record fields rely on the

characteristics of identified duplicate records. There were 38 groups of 2–4 case reports identified manually as suspected duplicates, available for this purpose.

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 were 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.

Hit-miss mixture models as described in Sect. 2.2 were fitted for the numerical record fields age (note as an aside the digit preference on 0 and 5) and date. Figure 4 shows empirical distributions in the WHO database for age and date together with empirical $f(d)$ functions. 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 (unbiased given $\mu_2 = 0$):

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

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:

$$\hat{a}_1 = 0.036 \quad \hat{a}_2 = 0.010 \quad \hat{b} = 0.186 \quad \hat{\sigma}_1 = 2.1 \quad \hat{\sigma}_2 = 32.9 \tag{22}$$

Table 4 Some parameters for the WHO hit-miss model

Record field	$\hat{a}$	$\hat{b}$	W_{jk}	Max W_{jj}	Min W_{jj}
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 (Unrelated death)	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)

The W_{jk} column lists mismatch weights and the two W_{jj} columns list the most extreme weights for matches in each record field

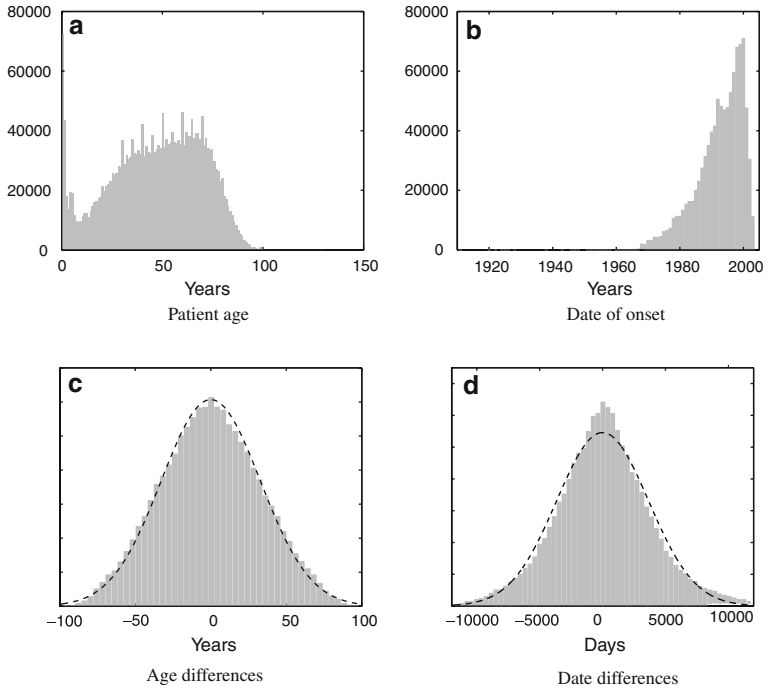

Fig. 4 Empirical distributions for Patient age and Date of onset on records in the WHO database, as well as empirical $f(d)$ distributions together with fitted normals

and for date:

$$\hat{a}_1 = 0.051 \quad \hat{a}_2 = 0.010 \quad \hat{b} = 0.229 \quad \hat{\sigma}_1 = 50.2 \quad \hat{\sigma}_2 = 3655 \tag{23}$$

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 in the fitted model (Fig. 5).

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 (because some groups of drug substances are often co-prescribed and certain drug substances cause particular reactions) so IC based compensation

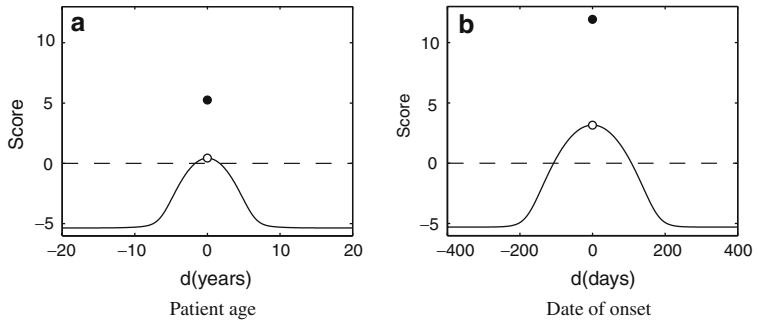

Fig. 5 Fitted hit-miss mixture model weight functions for Patient age and Date of onset, respectively. Note the discrete jump in the weight functions at $d = 0$

according to Sect. 2.3 was introduced for drug substances and ADR terms as one group.

3.3 A match score threshold

High match scores indicate likely duplicates, but in order to effectively discriminate duplicates from random matches we must set an appropriate threshold for manual review. Given the overall proportion of duplicates in the database together with match score distributions for duplicates and random matches respectively, Bayes formula can be used to calculate the false match rate at a given threshold. A key challenge is that both the proportion of duplicates and the two match score distributions are unknown. For record linkage applications, Belin and Rubin (1995) propose an approach where a mixture of two (transformed) normals is fitted to the overall distribution of match scores. The problem with such an approach for duplicate detection is the relatively small number of true duplicates compared to the number of unrelated record pairs. For the WHO database, the number of possible record pairs is in the order of 3 million squared, out of which at most a few hundred thousand record pairs will correspond to duplicates. As a simple but pragmatic alternative, we use the study of duplicate records in vaccine spontaneous reporting data (Nkanza and Walop 2004) as a proxy for the overall rate of duplicates and fit separate normal distributions to the observed match scores for the 38 duplicates in the training data (see Sect. 3.2) and a sample of 10,000 randomly matched pairs of records in the WHO database. The estimated means and standard errors were, for the labelled duplicates:

$$\hat{\mu}_r = 42.96 \quad \hat{\sigma}_r = 15.73 \tag{24}$$

and for the random matches:

$$\hat{\mu}_u = -18.50 \quad \hat{\sigma}_u = 8.55 \tag{25}$$

Based on this, we use Bayes formula to compute the the estimated true match rate for a given match score s:

$$\hat{P}(\text{dup} \mid s) = \frac{\frac{0.05}{3 \cdot 10^6} \cdot \phi(s; \hat{\mu}_r, \hat{\sigma}_r)}{\frac{0.05}{3 \cdot 10^6} \cdot \phi(s; \hat{\mu}_r, \hat{\sigma}_r) + (1 - \frac{0.05}{3 \cdot 10^6}) \cdot \phi(s; \hat{\mu}_u, \hat{\sigma}_u)} \tag{26}$$

In order to obtain an estimated false match rate of below 0.05, the match score threshold is set at 37.6 since $\hat{P}(\text{dup} \mid 37.6) \approx 0.95$ according to (26). The assumed 5% rate of duplicates in the database does not have a very strong impact on the threshold: a 20% rate would give a threshold of 35.8, a 10% rate would give a threshold of 36.7 and a 1% rate would give a threshold of 39.6.

Record fields with missing data do not contribute to the match score and mismatches contribute negatively, so a report can never receive a higher match score with another report than with itself. As a consequence, we may discard from any discriminative duplicate detection analysis those reports that have self match scores of below the given threshold. Such unmatchable reports (with respect to the threshold) usually have large degrees of missing data or unusually low information content (a tendency to have the most common record field values).

Whereas pairs of duplicates generate at most a single match per pair, the number of possible pairwise matches from larger groups of duplicates increases exponentially with the size of the group. In order to estimate the actual number of duplicates (and produce a more user friendly output), we use transitive closure (single link partitional clustering) to transform the list of pairwise matches to a list of case report clusters. Each such cluster contains all case reports with a pairwise match to least one other cluster member.

4 Results

4.1 Duplicate detection for a given database record

The aim of our first study was to evaluate the performance of the extended hit-miss model in identifying the most likely duplicate for a given database record. The test data set consisted of the 38 groups of manually identified duplicates described in Sect. 3.2 and to ensure independence, only the two most recent records in each group were used. The most recent record was designated the template record and the second most recent record the test record. Each template record was scored against all other records within its block (see Sect. 3.1) in the entire WHO database to see what proportion of the test records received the highest match score with their template records. While the same case reports had been used in estimating the parameters of the hit-miss model to be evaluated, their only impact was on a, a_1 and a_2 (the proportion of misses in different record fields) so the risk for over-fitting should be small.

For 36 of the 38 (94.7%) template records, the test record received the highest match score. The two imperfectly recalled template records are displayed in

Table 5 The first imperfectly recalled template record together with its best matches according to the hit-miss model

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

The test record is marked with an asterisk

Table 6 The second imperfectly recalled template record together with its best matches according to the hit-miss model

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

The test record is marked with an asterisk

Tables 5 and 6 together with the best matches according to the hit-miss model. For the template record in Table 5, there are no strong matches at all. Two records received slightly higher match scores than the test record, but do not seem like less plausible duplicates. For the template record in Table 6, there are 3 strong matches (match scores ranging from 37.78 to 47.28) with database records other than the test record. While these may well be false positives, they could also be undetected duplicates: they match on most of the record fields and although some of the ADR terms differ, a more careful analysis shows that they generally 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.

4.2 Discriminant duplicate detection

The aim of our second study was to evaluate the performance of the hit-miss model in discriminating true duplicates from random matches based on the threshold derived in Sect. 3.3. The threshold had been optimised with respect to the set of labelled duplicates used in the first study, so a new test data set was required. Fortunately, Norway who is one of few countries that provide information on confirmed duplicates in their own data set, had in their last batch in 2004 labelled 19 case reports as confirmed duplicates. This allowed for an independent evaluation of the duplicate detection method, in a data subset where the number of unidentified duplicates was expected to be low. Match

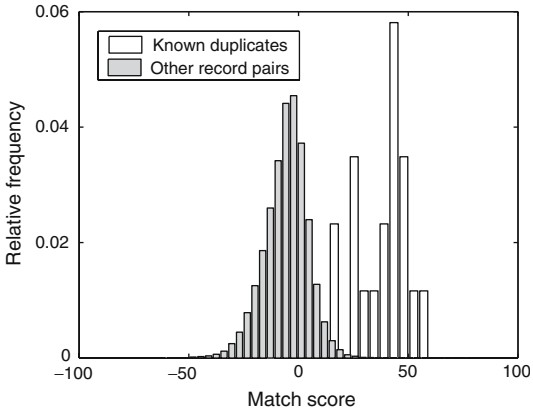

Fig. 6 Normalised match score distributions for known duplicates and other record pairs in the Norwegian batch

scores within blocks (see Sect. 3.1) were calculated for all case reports in the Norwegian database, and report pairs with match scores exceeding the 37.6 threshold were highlighted as suspected duplicates.

The total number of case reports in the Norwegian batch was 1559. The median match score for the 19 labelled duplicates was 41.8 and for randomly paired records within a common block −4.8. Fig. 6 displays the match score distributions for the two groups. All in all, 17 record pairs had match scores above 37.6 and out of these, 12 correspond to known duplicates and 5 to other record pairs. Thus, the recall of the algorithm was 63% (12 of the 19 labelled duplicates were highlighted) and the precision was 71% (12 of the 17 highlighted record pairs are among the labelled duplicates). However, the threshold of 37.6 was based on several assumptions, and following the discussion of precision-recall graphs by Bilenko and Mooney (2003b) Fig. 7 indicates how the precision and the recall varies with the threshold. To achieve the minimum total number of misclassifications, 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 Fig. 7, because the highest match score actually corresponds to a pair of records that were not labelled as duplicates. Table 7 lists this record pair together with the two other clusters of non-labelled case reports that were highlighted as suspected duplicates in the study. Table 8 lists the three labelled pairs of duplicates that received the lowest match scores in the study.

4.3 A database wide screen for duplicates

In order to study the feasibility of large scale duplicate detection in the WHO database, we carried out a screen for duplicates in the entire database (as of December 2004, excluding reports on vaccines and from clinical trials). The study was based on the same match score threshold as in the study of Norwegian data in Sect. 4.2 and transitive closure was used to translate the lists of

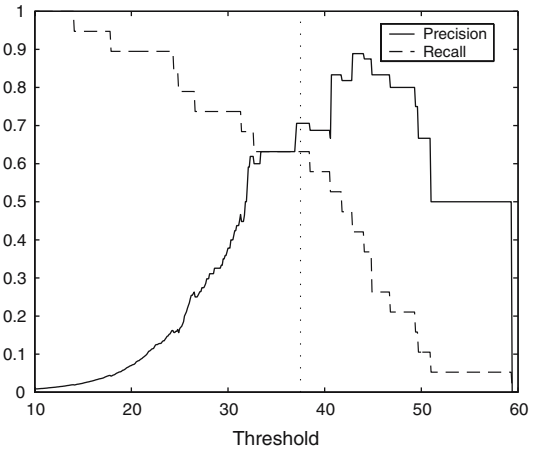

Fig. 7 Precision and recall as functions of the threshold, for the discriminant analysis of Norwegian data. The dotted line indicates the selected threshold

Table 7 Report pairs (and a triplet) highlighted in the Norwegian study that were not among the labelled duplicates

Onset date	Age	Gender	Country	Outcome	Drugs	ADR terms
2004-04-30	51	F	NOR	?	6 matched	0 matched + 2
2004-04-20	50	F	NOR	?	6 matched + 1	0 matched + 1
2003-02-02	57	M	NOR	?	2 matched	1 matched
2003-02-02	55	M	NOR	?	2 matched + 1	1 matched
2003-12-16	8	F	NOR	?	1 matched	1 matched
2003-12-16	18	F	NOR	?	1 matched	1 matched
2003-12-16	29	F	NOR	?	1 matched	1 matched

record pairs to groups of duplicates (see Sect. 3.3). More than 50,000 suspected duplicates were identified, which corresponds to about 1.8% of the evaluated data set. At the same time, more than 900,000 ($> 30\%$) case reports did not carry enough information to allow for any match at the selected threshold. Figure 8 displays a breakdown of all reports in the WHO database according to whether they are suspected duplicates, do not contain enough information to be reliably matched or do contain enough information and are not suspected duplicates. Figure 9 displays the same information for six individual countries (each with at least 5,000 reports in the database). Clearly, the variation between countries is large, with Finland, France and Romania being examples of countries with low proportions of suspected duplicates and few unmatchable reports. A more comprehensive breakdown by country for those with more than 5,000 reports in the data set is given in Fig. 10.

4.4 Computational requirements

The first two studies 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

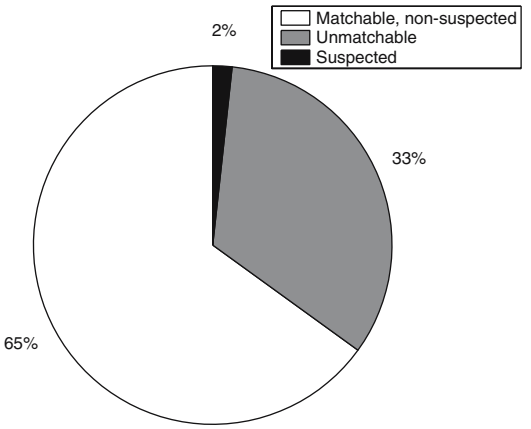

Fig. 8 Breakdown of reports for the entire WHO database

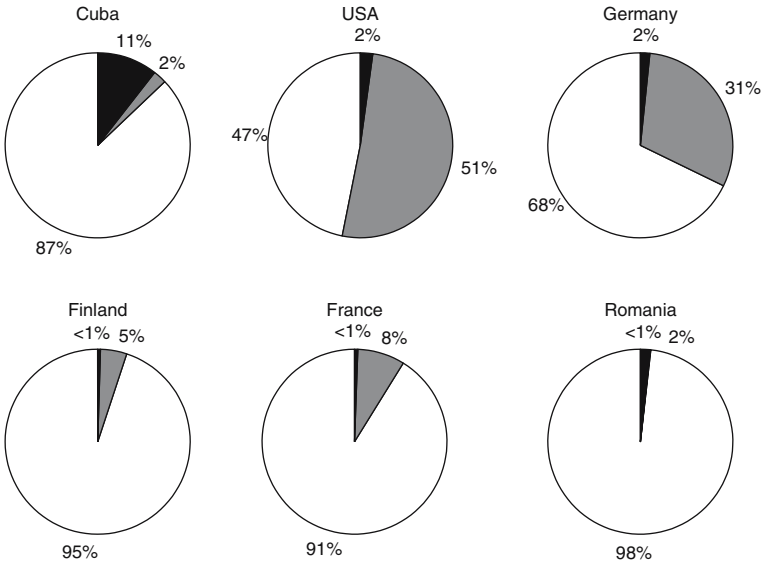

Fig. 9 Examples to illustrate the variation between countries (each with at least 5,000 reports in the database)

data structures reduced computing time and memory requirements so that the initial data extraction and model fitting required a total of 50 min for the entire WHO database. To score a single pair of database records took 6μs, and to score a database record against the rest of the data set took around 1 s (average block size in the order of 100,000 records). The screening for duplicates among the 1,559 record pairs in the Norwegian data subset took 27 s (with blocking). The database wide screen for all duplicates of all 3 million reports in the database required a total of around 100 h on a computer equipped with a dual 3.4 GHz Xeon processor and 3 GB of RAM.

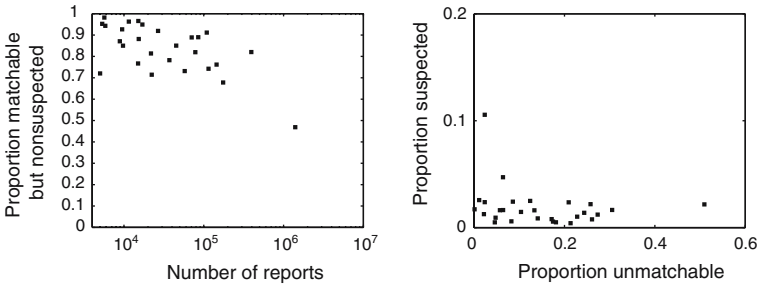

Fig. 10 Country variation (for those with more than 5,000 reports in the data set). To the left, the proportion of matchable but non-suspected case reports plotted against the total number of case reports. To the right, the proportion of suspected duplicates against the proportion of unmatchable reports

5 Discussion

The hit-miss model provides a rigorous framework for record matching, with very intuitive 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. It provides a prioritisation with respect to the chance that a given pair of records are duplicates, which allows the number of suspected duplicates highlighted in a particular study to be adjusted depending on the resources available for manual review. While penalising 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 but imperfect matches. Finally, because most of the hit-miss model parameters are determined by general properties of the entire data set, the risk of over-fitting the algorithms to training data is small. This is very important for the WHO database, where the amount of labelled training data is limited.

The results on WHO data are very promising. For case reports that were known to be duplicated, the hit-miss model reliably recalled the known duplicate (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 estimate for the probability that two records are duplicates. The 63% recall and 71% precision in Sect. 4.2 indicate that the hit-miss model identified the majority of labelled duplicates in the Norwegian data, while generating few false leads, which demonstrates its practical usefulness.

The hit-miss model did fail to highlight seven of the labelled duplicates in the Norwegian batch, but from Table 8 it is clear that these reports carry very little information: ages, outcomes and onset dates are missing on at least one of the records in each pair and while there are some 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

Table 8 The three lowest scoring report pairs among labelled duplicates in the Norwegian study

Onset date	Age	Gender	Country	Outcome	Drugs	ADR terms
?	79	F	NOR	?	1 matched	1 matched + 1
?	?	F	NOR	?	1 matched	1 matched + 1
2003-01-07	76	F	NOR	?	1 matched + 1	1 matched + 4
?	?	F	NOR	?	1 matched	1 matched + 1
?	43	F	NOR	?	2 matched + 2	0 matched + 7
?	?	F	NOR	?	2 matched	0 matched + 1

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 information cannot be compensated for with advanced algorithms. This emphasises the need for improved quality of case reports and in fact, the most critical data quality problem highlighted in the database wide screen reported on in Sect. 4.3 is not the 50,000 suspected duplicates (which can be removed upon confirmation), but the large number of reports that did not contain enough information to be reliably matched in the first place. The missing data problem must be addressed at data entry, but the total information content may be improved further by including additional record fields in the matching algorithm. One possibly informative record field for the WHO database which has not yet been used is the treatment start date. Treatment start dates are likely to be strongly correlated with ADR onset dates (and the compensation for correlated record fields is difficult to generalise to numerical variables), so the difference in days between the treatment start date and the ADR onset date may be a good choice of variable to add to the hit-miss model in the future.

Five record pairs highlighted in the Norwegian batch were not among the labelled duplicates (see Table 7). The first of these pairs received the highest match score in the entire study, but did not initially strike us as 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. What generated the unusually high match score is the simultaneous match on six different drug substances. These drug substances are not particularly commonly co-reported in general (the pairwise associations between them are relatively weak) which further strengthens the suspicion. In order to determine the true status of this pair of case reports, we contacted the Norwegian national centre who informed us that they are indeed labelled as duplicates in their data set: two different physicians at the same hospital have provided separate case reports for the same incident. The example 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 four other record pairs of unknown status that had been highlighted in the study: the record pair with the second highest match score is a likely but yet unconfirmed duplicate whereas the remaining three case reports are confirmed non-duplicates. However, these case reports

were provided by the same dentist, refer to the same drug-ADR combination and have the same listed onset date (possibly a data quality problem), which underlines the fact that the hit-miss model contrasts the hypothesis that two records relate to the exact same real world entity to the hypothesis that they are altogether unrelated. In reality, many record pairs, like this one, fall somewhere in between these two extremes. Clusters of similar case reports for different participants in the same clinical trial or for patients vaccinated simultaneously are another. With respect to duplicate detection, these are false matches, but in a different context the detection of related non-duplicate case reports will be very valuable (since they are considered less convincing evidence of a true problem than case reports provided independently). The Norwegian feedback indicates that the reported 71% precision in Sect. 4.2 is an under-estimate. The actual precision in the study 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 in the data set.

The hit-miss mixture model is a new approach to model 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 match: 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 data. There is a limit to how strongly negative the weight for a mismatch will get (see Fig. 5), so any large enough deviation is considered equally unlikely. An alternative model for dates which would be useful if typing errors were very common would be for year, month and day as separate variables. The disadvantage of such an 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. None of the record fields included in the hit-miss model for the WHO database contain free text, but by fitting hit-miss mixture models to string dissimilarity measures such as the edit distance or the vector-space cosine similarity, free text matching is possible within the hit-miss model framework.

The method to compensate for correlated record fields proposed in Sect. 2.3 allows for more robust record matching in the presence of non-independent categorical record fields. The compensation for pairwise associations reduces the risk for false positives due to matches on sets of associated record fields, but when there are higher order associations between record fields so that e.g. $P(j_3 \mid j_1, j_2)$ considerably exceeds all of $P(j_3)$, $P(j_3 \mid j_1)$ and $P(j_3 \mid j_2)$, the unexpectedness of a given set of matching events may still be over-estimated. In that case, an extension of the method in Sect. 2.3 to compensate for

interactions may be motivated. However, the resulting increase in model complexity would have to be balanced by a corresponding increase in the amount of relevant training data, and computationally efficient strategies for robust and automatic estimation of interaction terms would have to be defined. We believe that for this application the compensation for pairwise associations between drug substances and ADR terms is an appropriate compromise between model sophistication and usability.

The hit-miss model will be used routinely for duplicate detection 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 in ADR surveillance 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 first two studies in this article were retrospective in the sense that they evaluated performance based on already identified duplicates. We aim to follow up the results from the database wide screen in order to obtain prospective precision estimates and more insight into how the algorithm may be best applied in practice. The hit-miss model fitted to the WHO drug safety database in Sect. 3.2 should be useful for duplicate detection in other ADR data sets, provided they contain similar information. A more sophisticated approach would be to use the methods described in this paper to fit a hit-miss model directly for the data set of interest, since its properties may differ from those of the WHO database and additional record fields may be available. The latter approach may be useful for general record matching applications as well.

6 Conclusions

In this paper we have introduced two generalisations of the standard hit-miss model and demonstrated the usefulness of the extended 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.

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.

References

Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM (1998) A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 54:315–321

Belin T, Rubin D (1995) A method for calibrating false-match rates in record linkage. J Am Stat Assoc 90: 694–707

Bilenko M, Mooney RJ (2003a) Adaptive duplicate detection using learnable string similarity measures. In: KDD '03: proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, NY, USA, pp 39–48

Bilenko M, Mooney RJ (2003b) On evaluation and training-set construction for duplicate detection. In: Proceedings of the KDD-2003 workshop on data cleaning, record linkage and object consolidation, pp 7–12

Bortnichak EA, Wise RP, Salive ME, Tilson HH (2001) Proactive safety surveillance. Pharmacoepidemiol Drug Safety 10:191–196

Brinker AD, Beitz J (2002) Spontaneous reports of thrombocytopenia in association with quinine: clinical attributes and timing related to regulatory action. Am J Hematol 70:313–317

Copas J, Hilton F (1990) Record linkage: statistical models for matching computer records. J R Stat Soc: Sers A 153(3):287–320

De Veaux RD, Hand DJ (2005) How to lie with bad data. Stat Sci 20(3):231–238

Edwards IR (1997) Adverse drug reactions: finding the needle in the haystack. Br Med J 315(7107):500

Edwards IR (1999) Spontaneous reporting – of what? Clinical concerns about drugs. Br J Clin Pharmacol 48(2):138–141

Edwards IR, Aronson JK (2000) Adverse drug reactions: definitions, diagnosis and management. Lancet 356(9237):1255–1259

Evans SJW (2000) Pharmacovigilance: a science or fielding emergencies? Stat Med 19(23):3199–3209

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39(11):27–34

Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64:1183–1210

Hernández MA, Stolfo SJ (1998) Real-world data is dirty: data cleansing and the merge/purge problem. Data Min Knowl Discov 2(1):9–37

Jaro M (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J Am Stat Assoc 84:414–420

Kim WY, Choi B-J, Hong EK, Kim S-K, Lee D (2003) A taxonomy of dirty data. Data Min Knowl Discov 7(1):81–99

Lindquist M (2004) Data quality management in pharmacovigilance. Drug Safety 27(12):857–870

Monge AE, Elkan C (1997) An efficient domain-independent algorithm for detecting approximately duplicate database records. Research issues on data mining knowledge discovery, Tucson, AZ.

Newcombe HB, Kennedy JM (1962) Record linkage: making maximum use of the discriminating power of identifying information. Commun ACM 5(11):563–566

Nkanza JN, Walop W (2004) Vaccine associated adverse event surveillance (VAAES) and quality assurance. Drug Safety 27:951–952

Norén GN, Bate A, Orre R, Edwards IR (2006) Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 25(21): 3740–3757

Norén GN, Orre R, Bate A (2005) A hit-miss model for duplicate detection in the WHO drug safety database. In: KDD '05: proceeding of the 11th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, NY, USA, pp 459–468

Orre R, Lansner A, Bate A, Lindquist M (2000) Bayesian neural networks with confidence estimations applied to data mining. Comput Stat Data Anal 34:473–493

Rawlins MD (1988) Spontaneous reporting of adverse drug reactions. II: Uses. Br J Clin Pharmacol 1(26):7–11

Sarawagi S, Bhamidipaty A (2002) Interactive deduplication using active learning. In: KDD '02: proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining. ACM Press, New York, NY, USA, pp 269–278