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 all 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

SIGKDD '05, Aug 21-24, 2005, Chicago, IL, USA

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 practise. 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, 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 magnitude 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: sometimes different sources (health professionals, national authorities, different companies) provide separate case reports related to the same event; other times, there are mistakes in linking to the existing record, any follow-up case reports 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 based on 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 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 benefits. It imposes no strict criteria that a pair of records must fulfil in order to be highlighted as suspected duplicates, which is a useful property 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 with respect to the chance that a given pair of records are duplicates. This allows the number of record pairs highlighted for manual review to be varied depending on the available resources. 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 deviations. Furthermore, the reward for matching information varies depending on how common the matching event is, so that for example a match on a rare adverese event is considered stronger evidence that 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 duplicate records.

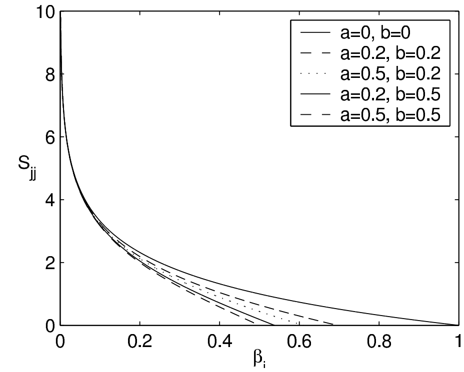


Figure 1: $W_{jj}(\beta_j)$ based on (8), for several values of a and b

2. METHODS

2.1 The hit-miss model

2.1.1 *The standard hit-miss model*

In the hit-miss model, the evidence in different record fields in favour and against two records being duplicates is added to provide a total match score for the record pair. Let $X = j$ and $Y = k$ denote the values in a certain record field on different database records and let p_{jk} denote the probability of observing this pair of values if the two records are duplicates. Similarly, let $p_j p_k$ denote the same probability under the assumption that the two records are unrelated. The contribution to the total match score from this observation (the weight for the record field) is equal to the log-likelihood ratio for these two hypotheses:

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

In the hit-miss model, each observation X is based on a true but unobserved event T. X is a random variable assumed to have been generated in a two-step process where the observed value is first classified as a miss (with probability a), a blank (with probability b) or a hit (with probability $1-a-b$), and then determined so that for a hit $X = T$, for a miss X is a random variable following the overall incidence of T and for a blank the value of X is missing.

Let $P(T = i) = \beta_i$ and let $P(X = j \mid T = i) = \alpha_{ji}$. The following holds generally under the assumption that the generation of X and Y is 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 $c = a(2-a-2b)$:

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

Outcomes	Probability	$f(d)$
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,R	$2a_2(1-a_1-a_2-b)$	$f(d)$
R,R	a_2^2	$f(d)$
D,R	$2a_1a_2$	approx $f(d)$

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

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

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

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

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 we would intuitively expect and 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, errors may occur for which small differences between true and observed values are more likely than large differences. 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 a patient may sometimes differ 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 difference is as likely as a small one. In order to handle both possibilities, we propose a hit-miss mixture model where each observation X is based on a true but unobserved event T. X is a random variable assumed to have been generated through a process that results in a limited deviation from T with probability a_1, an altogether random value with probability a_2, a blank with probability b and a hit with probability $1-a_1-a_2-b$. For a blank, the value of X is missing, for a hit, $X=T$, for a deviation X

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.

follows a $N(T,\sigma_1^2)$ distribution and for a random value X is independent of T but follows the same distribution.

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 distribution of T, such as for example random values and hits. The same is approximately true for deviations, under the assumption that $\text{var}(T) \gg \sigma_1^2$.

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

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

For unrelated records, d follows the more simple distribution:

$$p_u(d) = (1-b)^2\cdot f(d) \tag{10}$$

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. Like in the standard hit-miss model, single or double blanks receive weight 0.

$f(d)$ 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, we need to run an EM mixture identifier. 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 ofr 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 (the curse of dimensionality). As a compromise we propose the following approximation that accounts only for pairwise associations:

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

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, to compensate for pairwise correlations, we can simply subtract from the total match score (calculated under the regular hit-miss model) a sum of terms on the following form:

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

Shrinkage estimates for such log-ratios have previously been used as robust strength of association measures to find interesting associations in the WHO drug safety database [1, 15]. These are referred to as IC values and are defined as:

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

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], and this is the advantage of using IC values rather than raw observed-to-expected ratios. In order to provide more robust scoring of correlated record fields, we propose IC values be used to estimate $\log_2 \frac{\beta_{j_t}^*}{\beta_{j_t}}$ in (18). We do not account for negative correlations.

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

A generalised hit-miss model was fit to the WHO drug safety database based on the data available at the end of 2003, and a set of 38 manually identified groups of duplicate records.

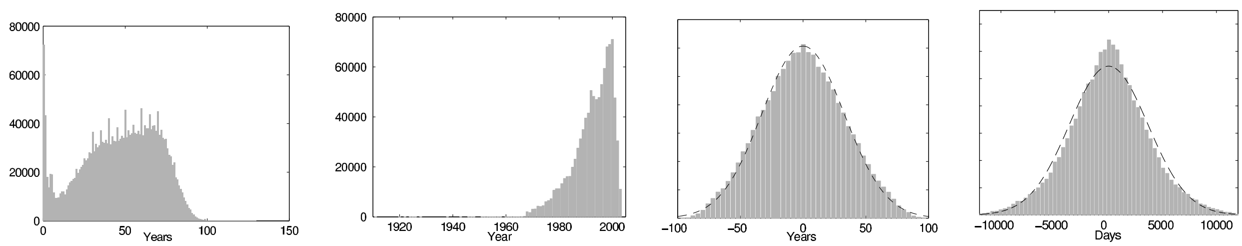
2.2.1 Implementation

In total, each WHO drug safety database record has 49(UPDATE!) different fields. These include both patient and administrative data, but the amount of information on each record is highly variable [1]. For the identification of possible duplicate records, the following record fields were considered the most relevant: date of incident, 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

Record field	Interpretation	Type	Missing data
DATE	Date of incident	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.

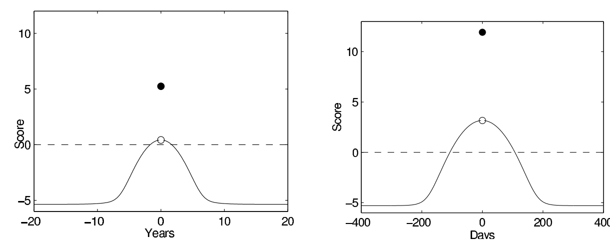
dates for different ADR terms; for those records, the earliest listed onset date was used. For the gender and outcome fields "-" had sometimes been used to denote missing values, and was thus re-encoded as such. Similarly, gender was sometimes listed as N/A which was also considered a missing value. 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 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 compared. 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).

2.2.2 *Training data*

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,

(REMOVE!?) Whereas the lack of identified duplicate



a. Age b. Date

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

records is a problem primarily in training the algorithms, the major challenge in evaluating performance is the large number of yet unidentified duplicates that still exist in the database.In order to properly evaluate the overall efficiency of duplicate detection algorithms in discriminant analysis, all duplicate records in the test data set must have been identified [4]. While this is not the case for the WHO drug safety database in general, it only affects the relative ranking experiment reported on in Section 3.1 insofar that the performance may be under-estimated. The relevance of the estimated performance of the hit-miss model for discriminant analysis as reported on in Section 3.2 is on the other hand conditional on the assumption that no significant amount of unidentified duplicates remain in the data subset under study.

2.2.3 *Model fitting*

Standard hit-miss models were fit to the gender, country and outcome record fields. Separate hit-miss models were fit for individual drug substances and ADR terms, but b and c was estimated for drug substances as a group and for ADR

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_{jj} values listed in columns 5 and 6 are the maximum and minimum weights for matches on different events in that particular record field.

terms as a group (c can be estimated based on (7) if $\sum \beta_i^2$ is 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 US origin are attributed much lower weights than matches on more rare events such as originating from a smaller country. 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:

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

and for date:

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

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 random 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 practise, 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.

2.2.4 *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 mixture model similar to that 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 but slightly skewed. 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 were the studies of duplicate records in vaccine spontaneous reporting data [14], which have 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).

2.2.5 *Experimental setup*

One experiment was carried out to evaluate the performance of the adapted hit-miss model in identifying the most likely duplicates for a given database record. The test data set consisted of the 38 groups of identified duplicates discussed in Section 2.2.2 and to avoid dependence between

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.

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 in the entire WHO database to see if any other pair received a higher match score with the template record than the test record.

Another experiment was carried out to evaluate the performance of the hit-miss model in discriminating duplicates from random record pairs based on the 37.5 threshold derived in Section 2.2.1. The test set used in the first experiment was not ideal for this purpose since these record pairs had been used to determine the threshold. However, Norway who is one of few countries that provide information on duplicate records had in their last batch in 2004 indicated 19 confirmed duplicates that allowed for an independent evaluation. Match scores were calculated for all record pairs within the same block. Those with match 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 it was the test records that received the highest match scores together with their template records, which was the case for 36 out of the 38 record pairs (94.7%). The two template records for which the test record was not successfully recalled are listed in Table 5 and Table 6 together with their 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 has been confirmed as one. Thus, while its performance was imperfect for this template record, the hit-miss model's prediction is in line with intuition. For the second difficult template record, there are strong matches (match scores ranging from 37.78 to 47.28) with three other records – none of which are known duplicates. While these 3 matches may be false positive, they could also be undetected duplicates: the records match on most of the record fields and although some of the ADR terms differ, a more careful analysis shows that all 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 Dicriminant 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 depends on the estimated 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. Precision might be anticipated to tend to 1 for higher threshold values, but this is not the case in Figure 5, because the highest match score actually corresponds to a record pair that are not known duplicates. Table 7 lists the three record pairs with highest match scores among other record pairs and Table 8 lists the three record pairs with lowest match scores

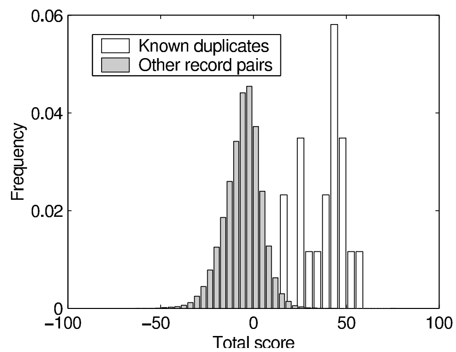


Figure 4: Match score distributions for known duplicates and other record pairs in the Norwegian batch, normalised in order to integrate to 1.

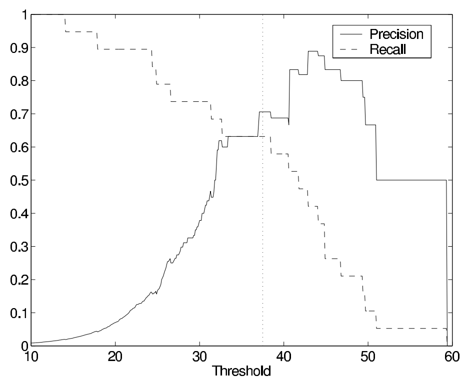


Figure 5: Precision and recall as functions of the threshold, for the discriminant analysis experiment on Norwegian data. The dotted line indicates the selected threshold.

among confirmed duplicates.

(DON'T KNOW HOW TO GET THIS IN!) The total number of false positives and false negatives is a possible summary performance measure. For the Norwegian data this measure is minimised for thresholds between 40.7 and 41.7 at 11 (2 false positives and 9 false negatives).

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 required to fit the hit-miss model 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 a second (average block size in the order of 100,000 records). The scoring for all record pairs the Norwegian data subset (1559 database records), after blocking, took 27 seconds.

4. DISCUSSION

For records that are known to have a duplicate, the hit-miss model reliably (94.7% accuracy) highlighted the corresponding record. However, only a small proportion of database records have duplicates so high ranked records are not necessarily duplicates. In order for the method to be truly effective at duplicate detection, it must instead provide an absolute estimate for the probability that two records are duplicates and it was the aim of the experiment in Section 3.2 to investigate this. The 63% recall and 71% precision in this experiment indicate that the hit-miss model identified the majority of known duplicates, while generating few false leads, and as such is a significant advance in duplicate detection for post-marketing drug safety data.The hit-miss model did fail to highlight 7 known duplicates in the Norwegian data and 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 for each pair, there are at least as many unmatched ones. The lack of data cannot be compensated for with advanced algorithms, which emphasises the need for improved quality of case reports for effective duplicate detection. The lowering of threshold required to highlight all these duplicates would lead to an unmanageable increase in the proportion of false leads, and we expect that any method would require non-anonymised data to be able to identify such duplicates. There are 5 highlighted record pairs that are not confirmed duplicates. One of these received the highest match score in the experiment (the top one in Table 7), but at a first glance this record pair does 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 has lead to the unusually high match score of 76.97. These drug substances are not particuarly commonly co-reported (the pairwise associations between them are weak) which strenghtens the evidence. In order to determine the true status of the record pair, we contacted the Norwegian national centre who were able to confirm that this is in fact a confirmed set of duplicates: two different physicians at the same hospital had provided separate case reports for the same incident. 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 reported to be confirmed non-duplicates. The last three made up a set of case reports that had been provided by the same dentist involving the same ADR terms in different patients. It is clear that case reports from the same individual will tend to be similar (particularly case reports from specialists who focus on certain types of patients) and may be difficult to distinguish from true duplicates. While not of primary interest in this context, it would be very valuable to be able discover individuals who send several similar reports, as this may affect the evaluation of a case series (independent case reports from different sources are considered stronger evidence). The Norwegian feedback indicates that the reported 71% precision in Section 3.2 is an under-estimate and that 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 in the data subset.

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

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 F	NOR NOR	? ?	1 matched, 0 unmatched	1 matched, 2 unmatched	24.36
2003-01-07 ?	76 ?	F F	NOR NOR	? ?	1 matched, 1 unmatched	1 matched, 3 unmatched	17.82
? ?	43 ?	F F	NOR NOR	? ?	2 matched, 2 unmatched	0 matched, 8 unmatched	14.05

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

The example with the previously unknown but now confirmed Norwegian duplicate illustrates the usefulness of the hit-miss model in real world duplicate detection. In particular, it shows that the hit-miss model may account for probabilistic aspects of data that may not be 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 practise.

The hit-miss mixture model presented in Section 2.1.2 is a new approach to handle discrepancies in numerical record fields. Like the standard hit-miss model, it is based on a rigorous probability model and provides intuitive results. For matches the weights depend on the precision: matches on full dates receive weights around 12.0, whereas 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. A pair of dates such as 1999-12-30 and 2000-01-02 contributes $+3.18$ to the match score, despite the superficial dissimilarity. Another intuitive property of the hit-miss mixture model is that there is a limit to how strongly negative the weight for a mismatch will get (see Figure 3): any large enough deviation is considered equally unlikely. An alternative approach which may be useful if typing errors are 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.

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 the results are followed up by manual review. Such a study should allow for more accurate precision estimates as well as more insight into how the algorithms may be best applied in practise.

The hit-miss model will be used routinely to detect likely duplicates in the WHO database – perhaps at data entry of all new case reports into the database, or automatically when a case series is selected for clinical review. As a complement, database wide screens for duplicates may be carried out but their computational requirements might necessitate changes to the method (DISCUSS!). The rate limiting step in any duplicate detection process is the manual review required to confirm or refute findings, so further testing will be necessary to set a threshold that 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 post-marketing drug safety data. Our results indicate that the hit-miss model may 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 the hit-miss model a strong candidate for other duplicate detection or 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*, 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.