Drug Saf (2014) 37:249–258
DOI 10.1007/s40264-014-0146-y

ORIGINAL RESEARCH ARTICLE

Performance of Probabilistic Method to Detect Duplicate Individual Case Safety Reports

Philip Michael Tregunno · Dorthe Bech Fink · Cristina Fernandez-Fernandez · Edurne Lázaro-Bengoa · G. Niklas Norén

Published online: 14 March 2014

Abstract

Background Individual case reports of suspected harm from medicines are fundamental for signal detection in postmarketing surveillance. Their effective analysis requires reliable data and one challenge is report duplication. These are multiple unlinked records describing the same suspected adverse drug reaction (ADR) in a particular patient. They distort statistical screening and can mislead clinical assessment. Many organisations rely on rule-based detection, but probabilistic record matching is an alternative.

Objectives The aim of this study was to evaluate probabilistic record matching for duplicate detection, and to characterise the main sources of duplicate reports within each data set.

Research Design vigiMatch™, a published probabilistic record matching algorithm, was applied to the WHO global individual case safety reports database, VigiBase®, for reports submitted between 2000 and 2010. Reported drugs, ADRs, patient age, sex, country of origin, and date of onset were considered in the matching. Suspected duplicates for the UK, Denmark, and Spain were reviewed and classified by the respective national centre. This included evaluation to determine whether confirmed duplicates had already been identified by in-house, rule-based screening. Furthermore, each confirmed duplicate was classified with respect to the likely source of duplication.

Measures For each country, the proportions of suspected duplicates classified as confirmed duplicates, likely duplicates, otherwise related, and unrelated were obtained. The proportions of confirmed or likely duplicates that were not previously known by the national organisation were determined, and variations in the rates of suspected duplicates across subsets of reports were characterised.

Results Overall, 2.5 % of the reports with sufficient information to be evaluated by vigiMatch were classified as suspected duplicates. The rates for the three countries considered in this study were 1.4 % (UK), 1.0 % (Denmark), and 0.7 % (Spain). Higher rates of suspected duplicates were observed for literature reports (11 %) and reports with fatal outcome (5 %), whereas a lower rate was observed for reports from consumers and non-health professionals (0.5 %). The predictive value for confirmed or likely duplicates among reports flagged as suspected duplicates by vigiMatch ranged from 86 % for the UK, to 64 % for Denmark and 33 % for Spain. The proportions of confirmed duplicates that were previously unknown to national centres ranged from 89 % for Spain, to 60 % for the UK and 38 % for Denmark, despite in-house duplicate detection processes in routine use. The proportion of unrelated cases among suspected duplicates were below 10 % for each national centre in the study.

P. M. Tregunno (✉)
Vigilance and Risk Management of Medicines Division (VRMM), Medicines and Healthcare Products Regulatory Agency (MHRA), 151 Buckingham Palace Road, London, UK
e-mail: phil.tregunno@mhra.gsi.gov.uk

D. B. Fink
Danish Health and Medicines Authority (DHMA), Copenhagen, Denmark

C. Fernandez-Fernandez · E. Lázaro-Bengoa
Agencia Espanola de Medicamentos y Productos Sanitarios (AEMPS), Madrid, Spain

G. N. Norén
Uppsala Monitoring Centre (UMC), WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden

G. N. Norén
Department of Mathematics, Stockholm University, Stockholm, Sweden

Conclusions Probabilistic record matching, as implemented in vigiMatch, achieved good predictive value for confirmed or likely duplicates in each data source. Most of the false positives corresponded to otherwise related reports; less than 10 % were altogether unrelated. A substantial proportion of the correctly identified duplicates had not previously been detected by national centre activity. On one hand, vigiMatch highlighted duplicates that had been missed by rule-based methods, and on the other hand its lower total number of suspected duplicates to review improved the accuracy of manual review.

1 Background

Effective pharmacovigilance requires trustworthy data [1]. Reports of suspected adverse drug reactions (ADR) remain a cornerstone of postmarketing surveillance, but duplicates may misguide their analysis [2–4]. Duplicates are separate and unlinked records that refer to one and the same case of a suspected ADR. They mislead assessors in their manual review and distort statistical screening. Duplicates do appear to come in clusters, as illustrated by an example where what appeared to be 20 reports on a very rare ADR boiled down to a single suspected case, reported multiple times [5].

Duplicates may stem from different reporters—a doctor and a nurse tending to the same patient, or the patient himself reporting in parallel to a health professional. They may also come from separate reports to different organisations such as regulatory agencies and pharmaceutical companies. This problem is particularly pronounced for case reports published in the literature for which reporting requirements force several organisations to enter the report in their systems, duplicating not only the report but also the effort. Finally, duplicates may result from unlinked follow-up reports or as an administrative by-product of errors in report transmission across and within organisations. The latter can be expected to increase in the future. The relative importance of each source of duplication has not been studied and is poorly understood.

Organisations with enough resources to perform individual case review for all reports may detect many suspected duplicates up front; this can be particularly effective in decentralised organisations where those who perform the review are close in time and space to the original reporter. Large organisations often rely on computerised duplicate detection, either home-grown or as part of a commercial software package, followed by subsequent manual review. The details of these duplicate detection algorithms are generally not published, but they tend to rely on heuristic rules such as: *if two reports match on the following fields, then they are suspected duplicates*.

The WHO global individual case safety reports database, VigiBase® [6], utilises vigiMatch™, a duplicate detection algorithm based on probabilistic record matching [2, 4]. vigiMatch does not enforce hard rules but scores each record field independently, adding up to provide an overall match score for the pair. For each record field, matching information is rewarded, and the reward is greater the rarer the matching events; mismatching information is penalised and the penalty is greater the rarer mismatches are in this record field, for known duplicates in the same database.

The aim of this study was to evaluate probabilistic record matching for duplicate detection and to characterise the main sources of duplicate reports within three national collections of individual case reports.

2 Methods

2.1 Data and Methods

The analysis was conducted in VigiBase®, with a special focus on suspected duplicates originating from the UK, Denmark, and Spain. Altogether, VigiBase® contains over 8 million reports of suspected harm from medicines, from 112 member countries across the world. Reports collected by national pharmacovigilance centres are pooled in VigiBase® for the purpose of detecting emerging safety signals early, within the WHO Programme for International Drug Monitoring [7]. Suspected duplicates identified in VigiBase® for each of these countries were analysed against the current information in the respective national databases, to determine if the cases are truly duplicates (see Fig. 1). More information about the approach to analysis is included in Sect. 2.3.

Each organisation participating in the study already employs its own duplicate detection algorithm. All three centres rely on rule-based matching and the methods used by the Medicines and Healthcare products Regulatory Agency (MHRA) and Danish Health and Medicines Authority (DHMA) are very similar. A more detailed description of these algorithms can be found in Table 1.

2.2 vigiMatch

vigiMatch is a duplicate detection algorithm based on the hit-miss model for record matching [2, 4]. The hit-miss model is a likelihood-based approach to identify unexpectedly similar record pairs in large databases [8]. It computes a match score for each pair of records, where matching information is rewarded and mismatching information penalised. This match score reflects the probability that the two records relate to the same underlying entity or,

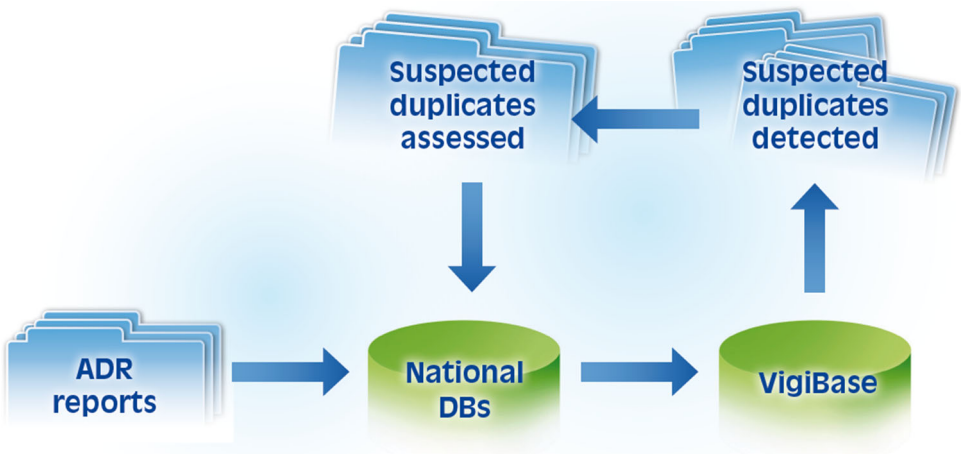

Fig. 1 Schematic overview of study: ADR reports are submitted to national centre DBs and transmitted to VigiBase®. vigiMatch identifies suspected duplicates in VigiBase® and the list of suspected duplicates originating in each respective country is sent to the national centre for detailed review and evaluation. *ADR* adverse drug reaction, *DBs* databases

in this setting, that they are duplicates. Record pairs with match scores that exceed a certain threshold are flagged as suspected duplicates. The threshold is derived from a comparison between the match scores of confirmed duplicates and of random record pairs in the database of interest. Formally, the hit-miss model score is a log-likelihood ratio for the hypothesis that the records relate to the same underlying entity (are duplicates) compared with the hypothesis that they are altogether unrelated. Reports with too little information cannot be highlighted by vigiMatch. vigiMatch ignores missing information and penalises mismatching information, so a report cannot receive a higher match score with another report than with itself (except in very special circumstances involving imprecise information for numerical fields). As a consequence, we label reports that fall below the match score threshold when compared against themselves as unmatchable and discard them from subsequent duplicate detection for computational efficiency.

Figure 2 provides a schematic illustration of how vigiMatch would apply to a pair of records. For each matching record field, a reward is added and for each mismatching field a penalty is detracted to form the total match score. vigiMatch exhibits characteristics as shown in Table 2 (for a detailed description, see Norén et al. [4]):

The implementation considered here is similar to that described in Norén et al. [4], but does not include 'Outcome' (since this is likely to differ between duplicates resulting from unlinked follow-up reports) and has been refitted to the current version of VigiBase®. It considers the following record fields: country of origin and patient sex (discrete); date of onset and patient age (numerical); drugs (WHO Drug Dictionary Enhanced™ substance level; suspected, concomitant, and interacting) and ADRs (WHO Adverse Reactions Terminology [WHO-ART] preferred terms) [binary vectors, with adjustment for correlations between drug pairs, ADR pairs, and drug-ADR pairs].

2.3 Empirical Evaluation

vigiMatch identified a list of suspected duplicates in VigiBase® for each national centre to review.

Each evaluated report pair was classified as:

- Confirmed duplicates
- Likely duplicates but as yet unconfirmed
- Not duplicates but otherwise related
- Unrelated
- Not in national dataset
- Other

Confirmed or likely duplicates were identified as:

- Previously known by national centre
- Previously unknown by national centre

The cause of duplication for each confirmed duplicate was classified as

- Reports of different origin
- Unlinked follow-up reports
- Errors in transmission
- Reports from multiple Marketing Authorisation Holders (MAHs)
- Other

The initial agreed scope of evaluation was all suspected duplicates from each country in VigiBase® between 1 January 2000 and 31 December 2010. However, due to the numbers of suspected duplicates identified, it was agreed to evaluate either all suspected duplicates, or the 100 most recent report pairs classified as suspected duplicates.

3 Results

3.1 Overview of Suspected Duplicates

Of the 3.7 million reports in VigiBase® entered between 1 January 2000 and 31 December 2010, 1.9 million (51 %)

Table 1 Comparison of organisations and their duplicate detection algorithms

Organisation	Fields used	Type of algorithm	References/ Publication
WHO Programme for International Drug Monitoring Total reports: 8,000,000 Annual growth: 500,000	• Country of origin • Drugs (substance level; suspected, concomitant, or interacting) • Adverse drug reactions (Preferred Term level) • Patient age • Patient sex • Reaction onset date	vigiMatch—probabilistic record linkage based on the hit-miss model (likelihood ratio-based)	[2, 4]
MHRA: Total reports: 700,000 Annual growth: 27,000	Country of origin Suspect drug (substance level) Patient details: Initials Age Gender Reporter name Safety report number Company Number/group Commit date (within last year)	Variable Matching (for healthcare professional/patient reports: suspect drug plus one of patient/reporter details; for industry reports, the above plus matching safety report IDs/ company numbers)	None
DHMA Total reports: 70,000 Annual growth: 6,000	Country of origin Suspect drug (substance level) Patient details: Initials Age Gender Reporter name Safety report number Company Number/group Commit date (within last year)	Variable matching (for healthcare professional/patient reports: suspect drug plus one of patient/reporter details; for industry reports, the above plus matching safety report IDs/ company numbers)	None
AEMPS Total reports: 200,000 Annual growth: 16,000	General algorithm: Autonomous Community Suspected drug (substance level) Adverse drug reaction MedDRA term (Preferred Term level) Start date of adverse drug reaction (year or ±6 months) Patient details: Gender Age (±10 units or = age group) Algorithm for industry's cases: Worldwide identification number (A.1.10) Safety report ID (A.1.0.1.) Duplicate (A.1.11.2.)	For cases received directly by healthcare professionals or patient, in addition to the automatic detection algorithm, pharmacovigilance centres perform a query in the database to detect possible duplicates For industry, there are three steps: • First step: algorithm for industry's cases (A.1.10), (A.1.0.1.) or (A.1.11.2) • Second step: general algorithm • Third step: Ad hoc queries. For example, cases from literature articles	None

MHRA Medicines and Healthcare products Regulatory Agency, *DHMA* Danish Health and Medicines Authority, *MedDRA* Medical Dictionary for Regulatory Activities

carry sufficient information to allow a suspected duplicate to be detected with vigiMatch. A report which when scored against itself does not attain a match score above the threshold is not informative enough to be matched against other reports and can be excluded from the duplicate screen to improve computational efficiency [2, 4]. Lower-than-

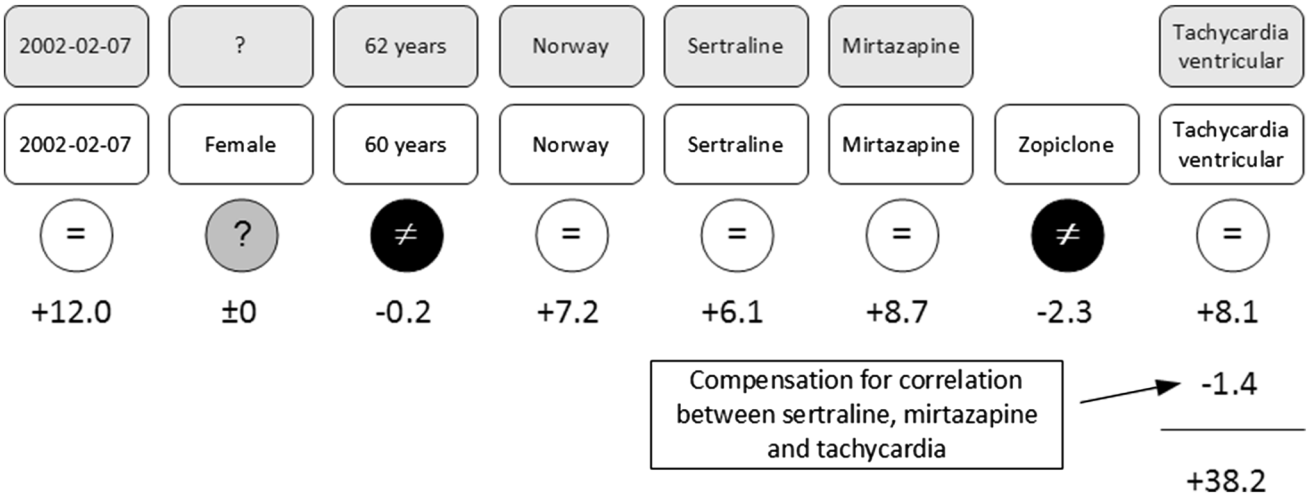

Fig. 2 Schematic illustration of hit-miss model scoring

Table 2 Characteristics exhibited by vigiMatch

Balances amount and agreement of information	The match score is driven by the *amount* of matching information. A pair of informative records with some mismatches may receive a higher score than a pair of identical records with minimal information
Derives from data	All parameters of the model are derived from data. The user decides which record fields to include, whether to treat them as discrete or numerical, and which correlations between record fields to adjust for. Most of the model parameters are estimated from general characteristics of the database, such as the proportion of records with missing information on patient sex, or the proportion of reports from France. The parameters related to penalties for mismatching information and the threshold for suspected duplicates are derived from confirmed duplicates
Rewards matches	Matches are always rewarded and the magnitude of the reward is determined by the relative frequency of the matched value in the database; the rarer the value, the higher the reward. Thus, a match on patient sex would yield a modest reward, whereas a match on date of onset would yield a substantial reward
Penalises mismatches	Mismatches are always penalised and the magnitude of the penalty is determined by how common mismatches are in that record field for pairs of confirmed duplicates in training data
Considers two extreme hypotheses	The hit-miss model considers two distinct hypotheses: that the two record pairs relate to the same underlying entity or that they are altogether unrelated. Record pairs that are related to one another but not true duplicates (in the case of adverse drug reaction reports they may refer to medically distinct adverse reactions for the same patient) may receive high match scores as they exhibit characteristics closer to what you would expect for a pair of duplicates than for independent reports
Allows for near-matches in numerical fields	The hit-miss mixture model extension for numerical record fields [2, 4] distinguishes mismatches based on the absolute deviation (the difference between two patient ages or dates of onset, for example). Close mismatches can be rewarded, but the reward for an exact match tends to be substantially higher
Can adjust for correlated record fields	The extended hit-miss model can compensate for correlated record fields. For individual case reports, adjustment for correlated drugs and adverse reactions reduces the match score for record pairs with a large number of matching drugs and adverse reactions that are often reported together (because the drugs tend to be used together or because the adverse reaction is associated with the drug or indication for treatment)

expected proportions of reports with sufficient information were observed for reports from the US (33 %), as well as for reports from lawyers (20 %), consumers/non-health professionals (26 %), and other health professionals (42 %), and for reports with a single reported drug (38 %).[1] Higher-than-expected proportions of reports with sufficient information were observed for a number of countries, as listed in Table 3, as well as for reports from clinical trials (89 %), special monitoring (89 %), and for reports from physicians (66 %).

In total, 48,000 clusters of suspected duplicates were detected, corresponding to 2.5 % of the reports with sufficient information. Significantly higher proportions of suspected duplicates were observed for reports from the literature (11 %), with fatal outcome (5.2 %), from other health professionals (4.9 %), and from studies (3.3 %).

[1] Matching drugs tend to be highly rewarded by vigiMatch, so the more drugs are listed on a report, the more likely they are to receive a sufficiently high score when matched against themselves.

Table 3 Proportions of reports with sufficient information for analysis by vigiMatch

Country	Percentage with sufficient information for analysis by vigiMatch
Japan	92
Sweden	89
Thailand	88
Malaysia	87
Cuba	84
France	84
New Zealand	83
Norway	82
Switzerland	81
Spain	78
Mexico	75
Germany	74
UK	73
The Netherlands	68
Australia	67

Countries with higher-than-average proportions of suspected duplicates include the Czech republic (15 %), Austria (15 %), Korea (9.2 %), and Switzerland (4.7 %). Lower proportions of suspected duplicates were observed for New Zealand (0.7 %), Spain (0.7 %), Japan (0.8 %), France (0.9 %), The Netherlands (0.9 %), Australia (1.0 %), and the UK (1.4 %). A lower proportion of suspected duplicates was also observed for reports from consumers/non-health professionals (0.5 %).

For the UK, the duplicate detection screen identified 1,862 suspected duplicates. This represents 1.4 % of the 140,000 VigiBase® reports from the UK with sufficient information in this time period. Higher proportions of suspected duplicates were observed for reports with fatal outcome (2.5 %), reports from the literature (9.3 %), and from studies (6.4 %).

For Denmark, the duplicate detection screen identified 136 suspected duplicates. This represents 1.0 % of the 13,000 VigiBase® reports from Denmark with sufficient information in this time period. Higher proportions of suspected duplicates were observed for reports from the literature (11 %) and from other health professionals (2.6 %).

For Spain, the duplicate detection screen identified 532 suspected duplicates. This represents 0.7 % of the 76,000 VigiBase® reports from Spain with sufficient information in this time period. Higher proportions of suspected duplicates were observed for reports from the literature (3.1 %) and from special monitoring (1.5 %).

The databases used for the analysis differ vastly in size and rate of growth, as can be seen in Table 1. VigiBase® comprises over 8 million reports from across the world. Out of the national databases considered in this study, the MHRA database is the largest, containing approximately 700,000 cases since its conception in 1963. By contrast, the DHMA and Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) datasets comprise approximately 70,000 and 200,000 reports, respectively.

3.2 Empirical Evaluation of vigiMatch

The 100 most recent clusters of suspected duplicates from the UK were evaluated as part of the study, while all 80 clusters from Denmark and all 276 clusters from Spain were evaluated. As shown in Fig. 3, the predictive value for confirmed or likely duplicates among reports flagged as suspected duplicates by vigiMatch ranged from 86 % for the UK, to 64 % for Denmark, and 33 % for Spain. Of these, 60 % from the UK, 38 % from Denmark, and 89 % from Spain were previously unknown to the national centre despite national duplicate detection processes. In 4 % of the clusters from the UK and in 1 % of those from both Denmark and Spain, evaluators considered that the cases were likely to be duplicates, although there was not sufficient information available to confirm their status. Fifty-three percent of the reports from Spain, 31 % of those from Denmark, and 11 % from the UK were classified as non-duplicates but related in another way. This included reports of different reactions for the same patient, reports for different patients in the same study, reports from the same health professional for different patients, and reports related to twins or parent-child type reactions. The proportion of all reports for the country that were classified as 'otherwise related' was similar for Spain ($0.7\ \% \cdot 0.53 = 0.37\ \%$) and Denmark ($1.0\ \% \cdot 0.36 = 31\ \%$). However, for the UK, it was lower ($1.4\ \% \cdot 0.11 = 0.15\ \%$), a phenomenon that we have not been able to explain. Two clusters from Denmark (2 %) and 26 from Spain (9 %) were classified by evaluators as entirely unrelated. Two such examples are provided in Table 4, both related to vaccines. Three reports from the UK and 12 from Spain could no longer be identified in the national dataset.

3.3 Sources of Duplicates

An overview of the reasons for report duplication across the three countries is shown in Fig. 4. Twenty-six percent of the confirmed duplicates from UK, 63 % of those from Denmark, and 38 % of those from Spain were caused by the national centre receiving separate reports directly from independent sources. This included different healthcare professionals reporting the same case as well as patients themselves reporting ADRs that had also been reported by a healthcare professional. Sixteen percent of the confirmed

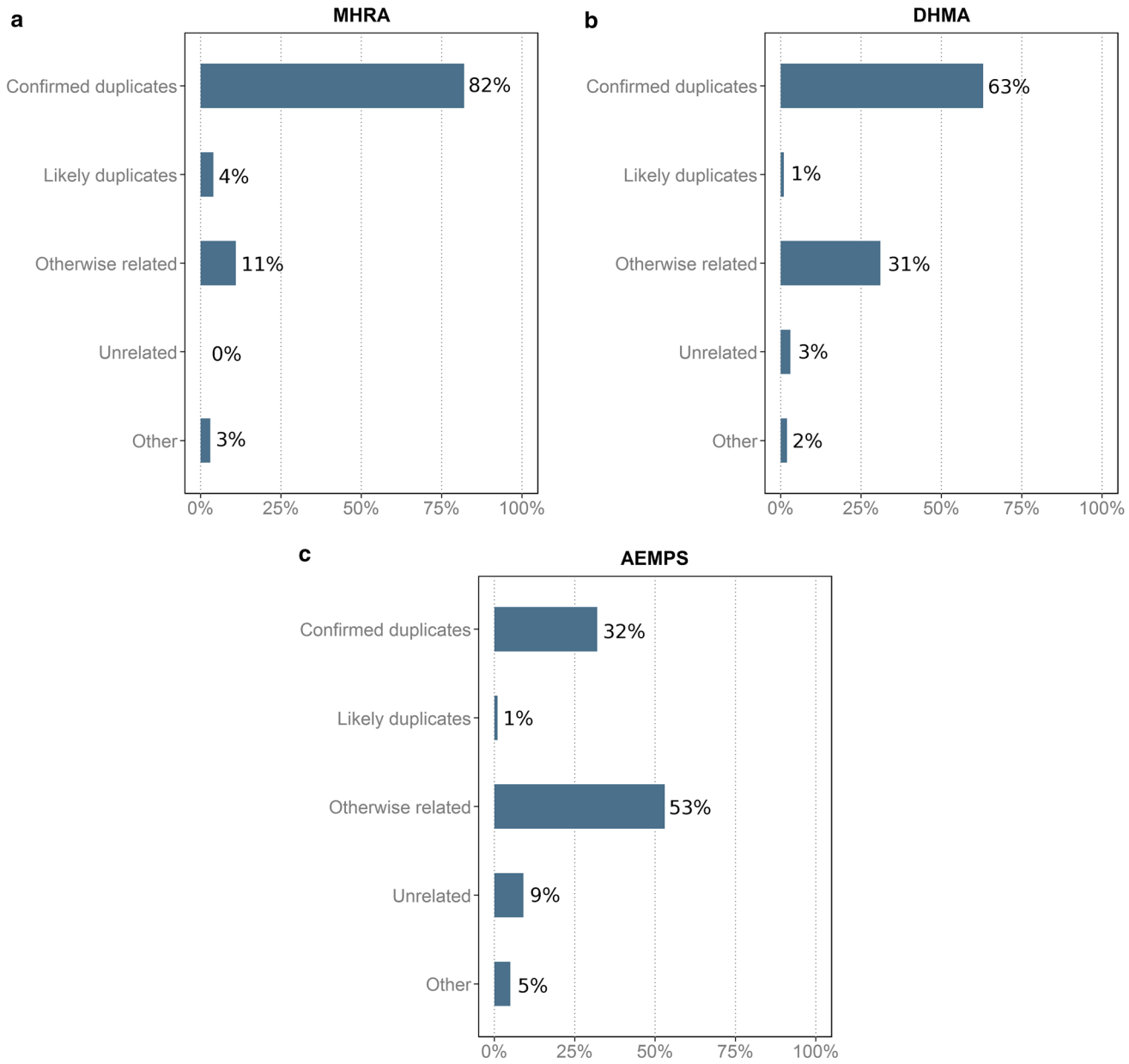

Fig. 3 True status of suspected duplicates for each respective country. **a** UK, in the MHRA database; **b** Denmark, in the DHMA database; **c** Spain, in the AEMPS database. *MHRA* Medicines and Healthcare products Regulatory Agency, *DHMA* Danish Health and Medicines Authority, *AEMPS* Agencia Española de Medicamentos y Productos Sanitarios

duplicates from the UK clusters and 32 % of those from Spain were the result of the national centre receiving follow-up cases that had not been linked to the original report. Notably, 35 % of the duplicates in VigiBase® from the UK were the result of transmission errors related to changes in reference numbers between case versions, in a database update. Some of these related to reports from MAHs with different reference numbers and others from transmission between the national database and VigiBase®; these were not duplicated in the MHRA database. Receipt of reports from multiple MAHs accounted for 15 % of the confirmed duplicates from the UK compared with 10 % for Denmark and 6 % for Spain.

3.4 Time for Evaluation

In addition to the results described above, evaluators additionally recorded the time taken for the duplicate analysis, not including time taken subsequently to merge the confirmed clusters. The MHRA's evaluation of 100 clusters took around 6 h, the DHMA's evaluation of 80 clusters took around 8 h, and the AEMPS' evaluation of 276 clusters

Table 4 Examples of report pairs flagged as suspected duplicates by vigiMatch but classified as unrelated in manual review

Country of origin	Patient sex	Patient age	Drugs	Adverse drug reactions	Date of onset
Denmark	Female	14 months	Mumps vaccine/rubella vaccine/measles vaccine	Fever Gait abnormal Rash	May 2002
Denmark	Female	17 months	Mumps vaccine/rubella vaccine/measles vaccine	Fever Gait abnormal	16 May 2002
Spain	Female	12 years	Hepatitis B vaccine	Eye abnormality Flushing Headache Skeletal pain	27 October 1998
Spain	Female	12 years	Hepatitis B vaccine	Flushing Headache Skeletal pain	27 October 1998

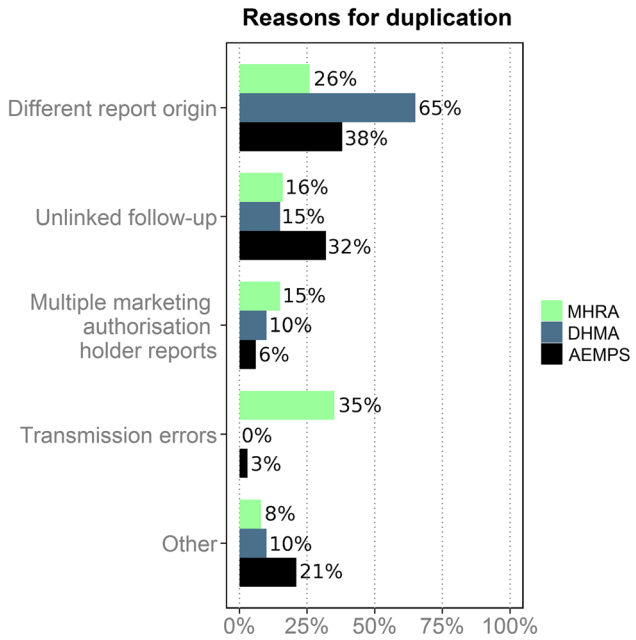

Fig. 4 Breakdown of reasons for duplication for the MHRA, DHMA and AEMPS. *MHRA* Medicines and Healthcare products Regulatory Agency, *DHMA* Danish Health and Medicines Authority, *AEMPS* Agencia Española de Medicamentos y Productos Sanitarios

took 18 h. This was as a result of the time taken to retrieve the cases and source documents from the database, compare information in different areas of the system, and consider other potentially related cases that had already been merged to the case or which had been flagged as suspected duplicates by the national duplicate detection system. In their evaluation, the AEMPS consulted each of the 17 regional centres of the Spanish pharmacovigilance system. For 17 % of the MHRA clusters, additional duplicates had been identified through national duplicate detection activities. Upon re-examination, only one of these additional duplicates were found to be present in VigiBase®, and represents a false negative for vigiMatch.

4 Discussion

Probabilistic record matching as implemented in vigiMatch achieved high predictive value for confirmed duplicates in each data source ranging from 82 % to 32 %. Pure false positives were rare: over 90 % of the cases evaluated in each country were related in some way, if not necessarily duplicates. A significant proportion of the confirmed duplicates had not been identified by the rule-based methods in use at the respective national centre. Assessors felt there were different reasons for this—on occasion, vigiMatch highlighted duplicates that would not be detected by the rule-based methods but, more commonly, vigiMatch's lower overall number of suspected duplicates gave assessors more time to review each pair, which led to improved accuracy. The rate of suspected duplicates from Spain was among the lowest of all countries in VigiBase®, and this may explain the higher proportion of otherwise related reports among Spanish record pairs highlighted by vigiMatch—many of the true duplicates had already been eliminated through efforts at the regional and national level, leaving a larger proportion of otherwise related reports among the suspected duplicates. Unfortunately, efforts to manually check cases before loading are resource-intensive and could no longer be feasible for Spain in the future.

The sources of duplication varied between countries but some similarities were observed. The most common cause was direct reporting of the same case from different reporters. For Denmark, this represented more than 60 % of all confirmed duplicates, and for the UK and Spain between 20 and 40 %. In contrast, multiple reports relating to the same case from different Market Authorisation Holders were more rare—around 10 % overall. This may reflect effective processes at the national centres to identify and merge such cases. Direct patient reporting began as a pilot in the UK in 2005 and was formalised in 2008, while

similar reporting schemes were initiated in Denmark in 2003 and in Spain at the beginning of 2013. These reports have added significant value to the signal detection process, with 24 % of signals of the MHRA having contributing reports from members of the public in 2010 (unpublished results). There was concern that direct patient reports would represent a source of additional duplicates, but in contrast the proportion of suspected duplicates among reports from patients in VigiBase® was lower than overall. On the other hand, a high rate of suspected duplicates was consistently observed for reports extracted from the scientific literature, a natural result of the responsibility of each pharmaceutical company and regulatory agency to identify and capture such reports. A high rate of suspected duplicates was also observed for fatal cases. It is believed that this is reflective of higher reporting rates for more serious events, where, for instance, a report can be received from a hospital doctor, hospital pharmacy, a general practitioner, and the patient themselves. A substantial number of duplicates from the UK were due to transmission errors and unlinked follow-up reports; the majority of these had arisen from transition to an E2B-based system in 2006 and subsequent transition to electronic submission for MAHs between 2006 and 2010. Duplicates arising from these scenarios were not necessarily duplicates at the national centre and these were exceptional circumstances. However, they do emphasize the need for care around database changes and subsequent re-transmission of affected cases.

There was a time lag between duplicate detection in VigiBase® and subsequent evaluation in the national centres. As a consequence, the level of duplication in VigiBase® for each country may be over-estimated. By the time of evaluation, some duplicates may have been highlighted and merged through the respective national processes. Although the lag is uncharacteristic of the duplicate detection systems used at national centres, significant duplication is caused by submission and re-transmission of the same case by multiple MAHs and regulators, prior to duplicate detection at each site. This is a result of timelines stipulated in the legislation for transmission of ADR reports, and emphasizes the need for swift and robust duplicate detection processes, and appropriate submission of nullification cases to organisations that have previously received the case.

The threshold used within vigiMatch for identification of a suspected duplicate is based on several assumptions that were not challenged during this study. The flexibility to adapt the threshold is an advantage of vigiMatch over rule-based methods in that it can be configured based upon the resources available for manual review. In our study, vigiMatch's false positive rate for unrelated cases was low, and the MHRA did identify one confirmed duplicate that existed in VigiBase® but had been missed by vigiMatch. In a previous evaluation against a set of reports with information on duplicate status, the algorithm's sensitivity was 63 %, and true duplicates that were not detected typically carried too little information to allow for a convincing match [4]. It would be valuable to explore the potential to lower the threshold to improve sensitivity. This would need to be balanced against the expected increase in the number of false positives. Effective duplicate detection requires informative reports. With too sparse details on each report, it is not possible to determine whether separate reports relate to the same suspected ADR, as illustrated by the fact that half of the reports are unmatchable by vigiMatch. In this context, national and international confidentiality laws applied by both MAHs and national centres in some member states can have a significant detrimental effect on duplicate detection efforts. Related to this, European Guidelines allow for the replacement of both reporter details and patient initials with terms such as 'PRIVACY', and if duplicate detection algorithms are not customised to account for this, this may result in false matches. vigiMatch incorporates a data preprocessing step where such snippets are marked as missing information.

From experience, it is understood that duplication can have a significant impact on disproportionality measures and can lead to false positive associations being investigated at the expense of true safety signals. A fundamental challenge is that duplicates are not evenly spread across the data: most reports have no duplicates and others have several. Unfortunately, the manual evaluation and elimination of suspected duplicates is extremely time-consuming and is not a viable option in many settings. An alternative approach is to exclude suspected duplicates from disproportionality analysis and adapt analytical software so that suspected duplicates can be highlighted to assessors in their clinical review. Further evaluation is required to determine the impact of this approach in real-life signal detection. An important aspect is the impact on disproportionality analysis of excluding otherwise related cases from screening. A benefit would be to help ensure the independence of reports, which is a fundamental assumption underlying the computation of confidence intervals for all disproportionality measures. Additionally, it reduces the undue impact of multiple reports from the same reporter, which should carry less weight than reports of the same quality from multiple independent sources. On the other hand, clusters of reports from the same reporter may be important for patient safety. They could result from appropriate and diligent reporting but could also reflect local risk patterns related to, for example, off-label use or medication errors. A better understanding of the reasons for otherwise related cases and their scientific implications for signal detection will help determine if this is a viable approach.

5 Limitations

Due to data privacy laws (which differ between EU member states), the number of data elements transmitted to the Uppsala Monitoring Centre are often not as rich as on the corresponding reports in the national databases. This, in turn, limits the potential of the algorithm in VigiBase®, and it is expected that there would be significant extra value in applying the method across a larger number of data elements. The implementation of vigiMatch directly on national data is a natural next step that would allow for direct comparison of duplicate detection methods when applied to the same collection of reports, utilising additional information such as patient initials.

6 Conclusions

Probabilistic record matching, as implemented in vigiMatch, achieved good predictive value for confirmed or likely duplicates in each data source. Most of the false positives corresponded to otherwise related reports; less than 10 % were altogether unrelated. A substantial proportion of the correctly identified duplicates had not previously been detected by national centre activity. On one hand, vigiMatch highlighted duplicates that had been missed by rule-based methods and, on the other hand, its lower total number of suspected duplicates to review improved the accuracy of manual review.

Acknowledgments The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi-protect.eu) which is a public-private partnership coordinated by the European Medicines Agency.

The PROTECT project has received support from the Innovative Medicine Initiative Joint Undertaking (www.imi.europa.eu) under Grant Agreement no. 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007–2013) and companies of the European Federation of Pharmaceutical Industries and Associations (EFPIA) in-kind contribution.

The authors would like to thank Johan Hopstadius, previously with the Uppsala Monitoring Centre, for contributions to the early phases of this project.

Conflicts of Interest G. Niklas Norén is an employee of the Uppsala Monitoring Centre who has developed and implemented the vigiMatch algorithm and may make it available as a commercial offering and/or as open source. Philp Michael Tregunno, Dorthe Bech Fink, Cristina Fernandez-Fernandez, and Edurne Lázaro-Bengoa have no conflicts of interest that are directly relevant to the content of this study.

References

1. Lindquist M. Data quality management in pharmacovigilance. Drug Saf. 2004;27(12):857–70.
2. Norén GN, Bate A, Orre R. A hit-miss model for duplicate detection in the WHO drug safety database. In: KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. New York, USA: ACM; 2005. pp. 459–68. doi:10.1145/1081870.1081923
3. Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. Drug Saf. 2005;28(11):981–1007.
4. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;2007(14):305–28.
5. Hauben M, Reich L, DeMicco J, Kim K. 'Extreme duplication' in the US FDA Adverse Events Reporting System database. Drug Saf. 2007;30(6):551–4.
6. Lindquist M. Vigibase, the WHO global ICSR database system: basic facts. Drug Inf J. 2008;42(5):409–19.
7. Olsson S. The role of the WHO programme on International Drug Monitoring in coordinating worldwide drug safety efforts. Drug Saf. 1998;19(1):1–10.
8. Copas JB, Hilton FJ. Record linkage: statistical models for matching computer records. J R Stat Soc Ser A Stat Soc. 1990;153(3):287–320.