Drug Saf (2017) 40:543–546
DOI 10.1007/s40264-017-0548-8

COMMENTARY

The Power of the Case Narrative - Can it be Brought to Bear on Duplicate Detection?

G. Niklas Norén[1]

Published online: 30 May 2017

1 What's in a Report?

Much of pharmacovigilance starts with the individual case report. The more detailed its description, the better it supports our causality assessment. The more explicitly it conveys the reasons that a reporter chose to write it up, the better we can appreciate its relevance and intention. The active contribution of data in support of reliable causality assessment is the greatest advantage of spontaneous reports over secondary use of claims or electronic health records, which are collected for other purposes.

The preparation of case data as structured information facilitates clinical review and enables statistical analyses. The latter might help organisations with large databases to highlight disproportional reporting of drugs and adverse effects for clinical review, and can identify key features of larger case series, such as unexpected patterns of time-to-onset.

Even so, all important features of a case cannot always be captured and conveyed as structured information. A primary objective in pharmacovigilance is to detect those risks that we did not know to look for. As a consequence, our standard terminologies will not always suffice to reflect the relevant nuances of clinical observations. Similarly, for signals that relate to the severity of the adverse event or its impact on the patient's quality of life, we might be at a loss to appreciate each individual's experience, based on the structured data alone.

This comment refers to the article available at doi:10.1007/s40264-017-0523-4.

✉ G. Niklas Norén
niklas.noren@who-umc.org

[1] Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden

In this issue of Drug Safety, Kreimeyer et al. [1] point to the opportunity that the power of case narratives might be brought to bear also on the challenge of duplicate detection.

2 It Never Rains but it Pours

Case report duplication is an important and growing obstacle to effective pharmacovigilance.

In the summer of 2002, when I first encountered the field of pharmacovigilance, there were 2.8 million reports in VigiBase, the WHO global database of individual case safety reports. These had been collected over 35 years, from 1968 to 2002. At present, VigiBase holds more than 14 million reports, with an annual growth of nearly 2 million, and its 2.8 million most recent reports have come in over the past 16 months.

We are becoming better at collecting and sharing information on suspected adverse drug reactions. This comes with both opportunities and challenges. It does seem that report duplication is not evenly spread—most reports have no duplicates at all, but others have several. Such report multiplication challenges the integrity of our case series and risks the reliability of both manual review and statistical signal detection.

3 Deduplication Design Decisions

There is a disappointing lack of published research on how to identify and account for duplicates among individual case reports, but the paper by Kreimeyer et al. [1] is an inspiring exception.

Perhaps the most novel and significant contribution of their research is the attempt to distil additional case information using text mining. Clearly, the more information that a report provides, the better positioned we are to assess if it is new and unique or a duplicate to an already reported adverse event—but this additional information can only improve automated duplicate detection to the extent that we can extract and use it in our algorithms. Kreimeyer et al. glean additional information that could ony be conveyed in free-text fields in their system (e.g. family history and medical history) or that could have been provided in structured format but was not (e.g. additional drugs and adverse events) [1].

While their results do not show a benefit of using text mining on duplicate detection, this absence of evidence is not evidence of its absence. More research is needed.

In general, I believe that enrichment of case details may well do more for duplicate detection than further methodological sophistication. Even so, I will highlight some differences and similarities between different choices of record matching method and point to some areas of uncertainty and further research.

The probabilistic record matching method used by Kreimeyer et al. (in combination with a rule-based approach) shares its heritage with vigiMatch, the duplicate detection method developed at the Uppsala Monitoring Centre and in use to monitor VigiBase for suspected duplicates [2]. Both are fundamentally based on the Fellegi–Sunter model, which contrasts the likelihood of different matching events (e.g. that two reports should list the same patient age) under two specific assumptions: (1) that the two reports refer to the same case and (2) that they are unrelated [3]. The corresponding log-likelihood ratios give match weights for each field, which can be added together for a total match score, under assumption of independence.

Kreimeyer et al. implement the Fellegi–Sunter likelihood ratio using an extension by DuVall et al. [4]. For fields such as age, country of origin, or date of birth, they base their match weights directly on the proportions of duplicates and non-duplicates that match on that field in training data. For fields that can contain multiple elements (such as the set of reported drugs), they measure a distance between two reports based on the proportion of the elements that they share, yielding a value between 0 (all elements in common) and 1 (no elements in common). They then look to training data to estimate the relative frequency with which distances for a specific field fall within a certain interval, for sets of known duplicates and non-duplicates, respectively. The ratios between these relative frequencies form the basis for their match weights.

vigiMatch implements the Fellegi–Sunter likelihood ratio based on the hit-miss model proposed by Copas and Hilton [5], with extensions to numerical fields and to correlated binary fields as described in Norén et al. [2]. The hit-miss model assumes a probabilistic model for how observed values are generated from the underlying true case, generating a hit (for which the observed value always matches the true value) with some probability and a miss (for which the observed value is random) by some other probability. These probabilities are estimated for each field, based on sets of known duplicates, and are combined with relative frequencies for different values of each field (such as country of origin = Peru) from the database as a whole, to yield Fellegi–Sunter likelihood ratios for different matching events.

A difference in principle between the two approaches is that Kreimeyer et al. use the empirical distance distributions directly, without assuming a specific underlying generative model for data. The stronger assumptions of the hit-miss model in return allow for the estimation of more precise match weights without access to larger sets of confirmed duplicates (the original vigiMatch implementation used no more than 38 pairs of confirmed duplicates). While Kreimeyer et al. attribute the same match weight to a matching country of origin, regardless of the country in question, vigiMatch attributes a higher match weight should two reports both be from Andorra (around 800 reports in VigiBase) than if they are from the Republic of Korea (around 900,000 reports in VigiBase). To estimate these match weights separately, directly from training data, would require massive amounts of training data, but vigiMatch instead estimates the general probability for a 'miss' on country of origin and combines this with the overall reporting rates from Andorra and the Republic of Korea to obtain their specific match weights.

Similarly, two reports that list the same set of drugs will receive the same match weight from Kreimeyer et al. (corresponding to distance 0), regardless of whether this set includes one or seven drugs. In vigiMatch, the corresponding match weight will depend on the number of matching drugs and how common each of them are in the database; as a consequence, reports that match on four out of five drugs are likely to receive a greater match weight with vigiMatch than reports that are identical but list only a single drug. Curiously, small non-zero distances do receive higher match weights than do zero distances also in Kreimeyer et al.'s implementations [1]. As noted by the authors, a likely explanation for this is that a majority of the pairs with zero distances are reports with single drugs (that match), which is weak evidence that the reports are duplicates. As a result, a pair of reports with five matching drugs and no mismatches will receive a lower match weight than a pair of reports with only four out of the same five drugs in common. On the other hand, one advantage of the distance-based approach by Kreimeyer et al. is that it limits the maximal impact that any one field may have on

the total match score. As a consequence, high match scores in their method are unlikely to be driven by matching information in a single field. In contrast, vigiMatch might flag reports with many drugs and adverse events in common as suspected duplicates even if they mismatch on all personal information such as patient age, sex, and date of onset, which can lead to false positives.

A challenge in common for the two approaches is how to handle the association between dates of onset and adverse events. Kreimeyer et al. treat dated and undated features separately, whereas vigiMatch considers only the earliest date of onset for any adverse event on the report. Neither approach is altogether satisfactory: for vigiMatch, the date of onset will be considered to mismatch if both reports list rash on March 22 but one of them additionally lists headache on March 7. For Kreimeyer et al., two reports that both list rash where only one lists March 22 will lead to two mismatches—one for the dated rash present only on the first report and one for the undated rash present only on the second report.

The differences between the two implementations of the Fellegi–Sunter model and their limitations highlighted above should be viewed in light of the over-arching design choice that they share: to use probabilistic record linkage. This is supported by research in the European public–private partnership IMI PROTECT, which showed a clear advantage over the more common rule-based approaches [6].

4 The Need for Speed

The challenges highlighted here and in the paper by Kreimeyer et al. should motivate us to pursue improved duplicate detection algorithms. What if we could attribute a match weight to the vaccination date, only when it differs from the adverse event onset date, since the two are clearly dependent? Or, perhaps time-to-onset is a better choice than vaccination date, since it should be independent of the date of onset? When we explore such improvements, we must bear computational tractability in mind. The computational complexity of duplicate detection is essentially quadratic: for 10,000 reports, there are in the order of 100 million possible pairwise comparisons and for 1 million reports there are 1000 billion (!). This can to some extent be alleviated with heuristics such as blocking or up-front exclusion of reports with too little information to allow for a reliable match [2], but it does mean that efficient database-wise duplicate detection will require pairwise comparisons that can be completed in (tiny) fractions of a second.

From this pespective, the use of text mining to extract additional features prior to duplicate detection is computationally advantageous, in that it can be done in a single pass over data prior to the actual record matching. Algorithms for direct text matching such as the CopyFind algorithm pointed to by Kreimeyer et al. would be an interesting complement. However, they need to be applied for each pair of duplicates within the record matching so are quadratic in computational complexity and would be challenging to incorporate in database-wide duplicate detection.

The focus of Kreimeyer et al. is duplicate detection of more limited scope, restricted to one single drug or vaccine at a time. The examples in their paper have only in the order of 1000 reports each. As they show, automated duplicate detection may bring value in this context too. For one thing, manual identification becomes challenging even for case series of moderate size: there are more than 400 possible pairs among 30 reports, and for 100 reports, there are 5000. Also, not all duplicates are easy for human assessors to detect, as illustrated by the examples of *false* false positives where record pairs first assumed to have been erroneously flagged by the algorithms, on closer inspection have turned out to be overlooked true duplicates [1, 2].

Back-end removal of suspected duplicates may facilitate clinical review, but will not reduce their distortion of statistical signal detection. In our experience, duplicate reports tend to have significant impact, in particular on the identification of more complex patterns, which are sensitive to lower case counts. As an example, duplicates were a major source of false positives in a recent effort of ours to detect harmful drug interactions in VigiBase (unpublished data). In this case, suspected duplicates had been identified and eliminated for the majority of our reports except those that had been received over the most recent few months. This lapse was enough to significantly disturb our screening, partly because single groups of duplicated reports would generate large numbers of false associations: a report in four copies listing seven drugs and four adverse events might generate 84 false leads ($\frac{7\times6}{2}$ drug pairs times the four adverse events), each with a case count inflated by three when the expected count would often be lower than one.

5 Patient Safety and Right to Privacy

The research by Kreimeyer et al. is important not only in advancing our methods for duplicate detection, but also in drawing attention to the need for high-quality case reports. In the absence of case details, the value of spontaneous reports dwindles. We may not be able to assess causality in the individual case and will struggle to determine which cases are unique, so as to determine the true sizes of our case series.

There is a rightful and growing focus on better protection of patient privacy. Going forward, we must work actively to safeguard the value that individual case reports bring to pharmacovigilance, without compromising patient privacy. Given the importance of case narratives, this will require research and development of methods to de-identify or de-sensitise individual case data. This is particularly challenging in the context of duplicate detection where the most useful information to determine if two cases are the same is often the most sensitive; this would include patient initials and birth dates, for which pseudonymisation may offer the best way forward.

In parallel, we must engage in policy discussions. One of the most simple and effective barriers against duplication is the use of worldwide unique identifiers for case reports. If such identifiers could no longer be shared between organisations for fear of compromising patient privacy, as a strict interpretation of the new European Data Protection legislation might suggest, then to my mind we are losing sight of why we collect these reports to begin with.

Compliance with Ethical Standards

Funding No sources of funding were used to assist in the preparation of this commentary.

Conflicts of interest G. Niklas Norén has no conflicts of interest that are directly relevant to the content of this commentary.

References

1. Kreimeyer K, Menschik D, Winiecki S, et al. Using probabilistic record linkage of structured and unstructured data to identify duplicate cases in spontaneous adverse event reporting systems. Drug Saf. 2017. doi:10.1007/s40264-017-0523-4.
2. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Min Knowl Discov. 2007;14(3):305–28.
3. Fellegi IP, Sunter AB. A theory for record linkage. J Am Stat Assoc. 1969;64(328):1183–210.
4. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators. J Biomed Inform. 2010;43(1):24–30.
5. 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.
6. Tregunno PM, Fink DB, Fernandez-Fernandez C, et al. Performance of probabilistic method to detect duplicate individual case safety reports. Drug Saf. 2014;37(4):249–58.