Representing Clinical Notes for Adverse Drug Event Detection

Aron Henriksson
Department of Computer and Systems Sciences
Stockholm University
Sweden
aronhen@dsv.su.se

Abstract

Electronic health records have emerged as a promising source of information for pharmacovigilance. Adverse drug events are, however, known to be heavily underreported, which makes it important to develop capabilities to detect such information automatically in clinical text. While machine learning offers possible solutions, it remains unclear how best to represent clinical notes in a manner conducive to learning high-performing predictive models. Here, 42 representations are explored in an empirical investigation using 27 real, clinical datasets, indicating that combining local and global (distributed) representations of words and named entities yields higher accuracy than using either in isolation. Subsequent analyses highlight the relative importance of various named entity classes for predicting adverse drug events.

1 Introduction

Electronic health records (EHRs) have emerged as a potentially valuable, and complementary, source of information for pharmacovigilance, which, as a result of the limitations of clinical trials – in terms of duration and sample size – needs to be carried out throughout the life-cycle of a drug to inform decisions about sustained use. Adverse drug events (ADEs), defined as undesired harms resulting from the use or misuse of a drug (Nebeker et al., 2004), are the most common iatrogenic injury, being responsible for around 3.7% of hospital admissions worldwide (Howard et al., 2007). The adverse effects of drugs cause suffering in patients and put an economic burden on healthcare – often unnecessarily, as ADEs are in many cases preventable (Hakkarainen et al., 2012).

A challenge for pharmacovigilance is that ADEs are heavily underreported (Hazell and Shakir, 2006), both in so-called spontaneous reporting systems, whereby reports of ADE cases are submitted voluntarily by patients and clinicians, and in EHRs, in which ADEs can be encoded by a set of diagnosis codes. To address the problem of underreporting, systems that can automatically detect ADEs in EHRs are potentially valuable, and much research has been conducted to that end (Harpaz et al., 2012). While many efforts have aimed at using machine learning for detecting ADEs on the basis of structured EHR data (Chazard et al., 2011; Zhao et al., 2014a; Zhao et al., 2014b; Zhao et al., 2015), attempts have also been made to exploit the more unstructured data in the form of clinical notes (Eriksson et al., 2013; LePendu et al., 2013). These have either relied on manually constructed rules and extensive dictionaries or on applying disproportionality methods¹ to counts of terms extracted from clinical notes. In a recent study (Henriksson et al., 2015a), information pertaining to ADEs – including named entities such as drugs and medical problems, as well as relations between them, i.e., whether one exists and whether it expresses, e.g., an indication or an ADE – were detected in clinical notes using machine learning; this approach, however, relies on the availability of data that has been manually labeled outside the clinical setting. There have also been efforts to combine information from the structured and unstructured sections of EHRs for ADE detection (Harpaz et al., 2010; Eriksson et al., 2014). In one of these (Henriksson et al., 2015b), heterogeneous types of clinical data, including free-text notes,

¹Disproportionality methods describe to what extent the co-occurrence frequency of two events deviates from what is expected (Suling and Pigeot, 2012).
were represented using distributional semantics, the use of which is also investigated in this study. In the previous study, however, many possible alternative ways of representing clinical notes were left unexplored. A more in-depth investigation is conducted in the present study, focusing on the representation of clinical notes for ADE detection.

In this study, ADE detection using clinical notes is approached as a binary classification task, in which the presence or absence of a particular ADE in a healthcare episode is to be determined; for this purpose, diagnosis codes assigned in the clinical setting are used as class labels. This raises the question of how best to represent clinical notes. There are certainly challenges involved in applying machine learning to high-dimensional and sparse data, which, as a result of prevalent misspellings and creative shorthand, clinical notes are a prime example of. These challenges will be considered when exploring possible representations of clinical notes.

2 Materials and Methods

This study explores 42 different ways of representing clinical notes and evaluates their effectiveness, in terms of classification accuracy, on the task of detecting the presence of an ADE in a healthcare episode. The use of both local and global (distributed) representations of words and named entities, as well as their combination, is investigated in an experiment using 27 ADE datasets, followed by a number of further analyses. Local representations are ones that do not incorporate any prior (semantic) knowledge of the similarity of token types, while global representations do, in this case by applying models of distributional semantics to a much larger corpus, resulting in word embeddings that are then exploited in the ADE detection task. While local representations are commonly employed for document classification, the use of global, distributed representations has been less thoroughly investigated, with a few exceptions (Sahlgren and Cöster, 2004; Henriksson et al., 2015b). Here, various types of local and global representations are compared and combined in an exploratory fashion.

2.1 Data Source

The 27 datasets were extracted from a Swedish EHR database (Dalianis et al., 2012), which contains health records over a two-year period from Karolinska University Hospital\(^2\). The learning task is to detect healthcare episodes that involve a certain ADE, i.e., in which an ADE-specific ICD-10 diagnosis code has been assigned. A healthcare episode is here defined based on the time interval between recorded activities for a patient, delimited by at least three days of inactivity. Each of the 27 datasets thus consists of healthcare episodes, where the positive examples have been assigned an ADE-related diagnosis code, and the negative examples are an equal number of randomly selected healthcare episodes in which that same code has not been assigned. The ADE-related diagnoses were selected on the basis of having been classified as indicating ADEs in a previous study (Stausberg and Hasford, 2011) and being sufficiently frequent (> 10) in the database. The datasets are described in Table 1. In addition to the labeled datasets, the entire two years of data is used for obtaining global, distributed representations of words. The notes, containing approximately 3M unique types (700M tokens), are preprocessed by using Stagger (Östling, 2013) for tokenization and lemmatization of Swedish text.

2.2 Data Representations

14 × 3 = 27 representations of clinical notes are explored. Each of the fourteen representations of words and/or named entities are weighted in three different ways. The local representations include the commonly employed unigrams, bigrams and trigrams, as well as their combination. In addition, a named entity recognition (NER) model trained on Swedish clinical text (Henriksson et al., 2015a) is applied to the healthcare episodes to extract mentions of the following named entity types: Finding, Disorder, Drug, Body Part and ADE Cue\(^3\). Local representations of identified named entities, without specifying type (denoted Terms), as well as a combination of unigrams and terms,

\(^2\)This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/834-31/5.

\(^3\)An ADE Cue corresponds to an expression that indicates the presence of an ADE without revealing its precise manifestation, e.g., side effect or drug-induced.
are explored.

In addition to the local representations, the use of global, distributed representations of words and terms is explored. Word embeddings are obtained using a recently introduced model of distributional semantics – see (Cohen and Widdows, 2009) for an overview – based on shallow neural networks with a single hidden layer: the skip-gram model (Mikolov et al., 2013) as implemented in word2vec. It was chosen for its ability to produce high-quality vector representations of words, outperforming traditional context-counting based methods on a range of tasks (Baroni et al., 2014). The algorithm obtains vector representations of the words in the training set by learning to predict nearby context words of each target word; the learned weights within the neural network are then used as vector representations. In a basic configuration, a symmetric context window size of 10 and a dimensionality of 200 is employed. Distributed representations of clinical notes are obtained by simply summing the vectors corresponding to the constituent token types; when representing notes by terms, the words that make up multiword terms are likewise summed. As it has been shown that improved performance can be obtained by combining various word representations (Henriksson et al., 2014), we also explore the use of distributed ensembles created by employing a number of different context window sizes: 6, 8, 10, 12, 14. The representations of healthcare episodes are then obtained by fusing the features from each distributional semantic space. The intuition behind this is that they will capture different aspects of the data. Both single distributed representations and ensembles thereof are used to model healthcare episodes as a combination of unigrams and terms.

Finally, combinations of local and global representations are explored: (1) combining local and global representations of unigrams and terms from a single semantic space, and (2) combining them from multiple semantic spaces. In all representations, the lowercase lemma of the tokens is used. The three weighting strategies are: (1) binary, (2) term frequency (TF), and (3) term frequency-inverse document frequency (TF-IDF). The binary representation corresponds to the so-called one-of- $K$ or one-hot encoding.

Table 1: Description of the 27 ADE datasets used in the study

| Dataset | Episodes | Types | Tokens | Types | Tokens | Types | Tokens | Types | Tokens |
|---------|----------|-------|--------|-------|--------|-------|--------|-------|--------|
| D64.2   | 416      | 58,455,264,271 | 432,235,2,380,942 | 895,864,2,121,106 | 18,204,187,885 |
| E27.3   | 34       | 10,990,142,365 | 55,922,127,740 | 78,222,113,136 | 2,729,11,907 |
| F11.0   | 76       | 13,906,285,629 | 73,688,257,154 | 107,216,228,707 | 3,532,21,448 |
| F11.2   | 308      | 35,340,1,118,139 | 93,847,2,642,271 | 144,340,212,272 | 4,204,18,262 |
| F13.0   | 120      | 16,759,264,555 | 73,413,237,228 | 106,560,211,607 | 3,546,19,140 |
| F13.2   | 32       | 48,919,1,118,139 | 25,674,38,814 | 31,226,88,854 | 1,362,3,755 |
| F15.0   | 46       | 10,849,139,693 | 51,897,123,224 | 72,081,110,391 | 2,438,9,776 |
| F15.1   | 256      | 30,098,812,312 | 193,693,729,918 | 304,491,647,691 | 8,896,61,548 |
| F15.2   | 256      | 30,098,812,312 | 193,693,729,918 | 304,491,647,691 | 8,896,61,548 |
| F19.0   | 122      | 18,257,341,225 | 100,890,307,757 | 152,834,274,353 | 4,671,23,638 |
| F19.1   | 74       | 14,279,216,583 | 75,268,194,519 | 109,675,172,500 | 3,457,16,504 |
| F19.2   | 288      | 34,925,992,236 | 229,330,891,735 | 404,862,791,349 | 10,645,76,797 |
| F19.9   | 68       | 14,938,221,480 | 78,506,198,688 | 112,658,175,942 | 3,571,16,205 |
| G24.0   | 28       | 11,293,125,342 | 57,711,112,935 | 81,196,100,542 | 2,897,9,454 |
| G62.0   | 20       | 5,121,44,776 | 19,564,40,609 | 23,359,36,449 | 1,027,3,350 |
| I95.2   | 70       | 13,321,179,622 | 69,127,161,505 | 99,732,143,442 | 3,336,14,651 |
| L27.0   | 274      | 41,669,1,394,815 | 281,601,1,255,535 | 513,539,1,116,442 | 12,908,109,896 |
| L27.1   | 78       | 15,495,291,266 | 84,668,261,351 | 125,632,231,491 | 4,118,23,824 |
| N14.1   | 28       | 10,383,101,969 | 49,864,92,098 | 67,286,82,247 | 2,585,7,979 |
| O35.5   | 128      | 11,810,145,344 | 57,172,131,075 | 79,992,116,826 | 2,745,10,313 |
| T59.9   | 63,551   | 57,773 | 26,831,51,763 | 34,017,45,766 | 1,474,4,499 |
| T78.2   | 102      | 15,272,236,489 | 80,533,212,753 | 118,274,189,062 | 3,845,19,461 |
| T78.3   | 266      | 26,716,503,385 | 161,366,451,831 | 265,928,400,439 | 7,912,43,832 |
| T78.4   | 1520     | 56,244,1,950,200 | 396,818,1,752,142 | 783,547,1,555,017 | 18,415,167,620 |
| T80.8   | 732      | 48,299,2,053,152 | 349,030,1,845,434 | 698,814,1,638,072 | 16,247,169,391 |
| T88.6   | 96       | 17,453,280,652 | 96,546,252,705 | 145,766,224,818 | 4,714,23,191 |
| T88.7   | 564      | 51,922,1,422,450 | 357,484,1,600,899 | 680,750,1,422,450 | 16,738,138,899 |

When representing notes by terms, the words that make up multiword terms are likewise summed. As it has been shown that improved performance can be obtained by combining various word representations (Henriksson et al., 2014), we also explore the use of distributed ensembles created by employing a number of different context window sizes: 6, 8, 10, 12, 14. The representations of healthcare episodes are then obtained by fusing the features from each distributional semantic space. The intuition behind this is that they will capture different aspects of the data. Both single distributed representations and ensembles thereof are used to model healthcare episodes as a combination of unigrams and terms.

Finally, combinations of local and global representations are explored: (1) combining local and global representations of unigrams and terms from a single semantic space, and (2) combining them from multiple semantic spaces. In all representations, the lowercase lemma of the tokens is used. The three weighting strategies are: (1) binary, (2) term frequency (TF), and (3) term frequency-inverse document frequency (TF-IDF). The binary representation corresponds to the so-called one-of- $K$ or one-hot encoding,
indicating the presence or absence of a feature; TF corresponds to the bag-of-words representations; finally, TF-IDF is the product of TF in a particular document and the term’s IDF. It thus gives less weight to common terms with little discriminative value.

### 2.3 Experimental Setup

The main experiment involves a comparison of the 42 representations and their impact on classification accuracy. Here, the random forest algorithm (Breiman, 2001) is used due to its reputation of achieving high accuracy, its ability to handle high-dimensional data, as well as the possibility of obtaining estimates of variable importance. The algorithm constructs an ensemble of decision trees, which together vote for what class label to assign to an example. Each tree in the forest is built from a bootstrap replicate of the original instances, while a subset of all features is sampled at each node when building the tree. This procedure is intended to increase diversity among the trees. When the number of trees in the forest increases, the probability that a majority of trees makes an error decreases, given that the trees perform better than random and that the errors are made independently. Although this can only be guaranteed in theory, the algorithm has often been shown in practice to result in state-of-the-art predictive performance. In this study, we use random forest with 500 trees, while \( \sqrt{n} \) of all available \( n \) features are inspected at each node.

Using the terms representation, a follow-up analysis is conducted to gain insight into which (types of) terms are most useful in the classification task. Variable importance can be estimated in different ways (Breiman, 2001). Here, Gini importance is used as the variable importance metric, where high Gini importance means that a feature plays a greater role in splitting the data into the defined classes. A Gini importance of zero indicates that a feature is considered useless or is never selected to build any tree. We inspect the twenty most important features, averaged over datasets, but we also calculate the average rank of terms of various lengths and named entity classes to understand which types of terms are more informative. Finally, the frequency of various named entity types across the two classes is analyzed in an attempt to identify potentially important differences.

Models are built and evaluated using ten iterations of stratified 10-fold cross validation. For testing the statistical significance of observed differences between the various representations, the Friedman test, as suggested in (Garcia and Herrera, 2008), is employed, where the null hypothesis is that the methods perform equally well.

### 3 Results

The accuracy scores, averaged over the 27 datasets, produced with the various data representations are shown in Table 2. A Friedman test rejects the null hypothesis that the various representations perform equally well \( (p < 0.0001) \). Of the three weighting strategies, the binary strategy performs almost invariably better than the TF and TF-IDF strategies. When comparing the ngram representations, unigrams perform considerably better than bigrams and trigrams, while their combination is plausibly negatively affected by the latter two. Using only extracted terms performs slightly better than using all unigrams or a combination of unigrams and terms, albeit the differences are small. The global, distributed representations only outperform the local representations when multiple semantic spaces are used in an ensemble. Moreover, all ensembles outperform their single-model counterparts. The best predictive performance is obtained when combining local and global representations – in a semantic space ensemble – of unigrams and terms, yielding an accuracy of 83.89%.

The twenty most important term features are listed in Table 3. All of these are names of drugs, findings and disorders. Some of the drugs are known to cause ADEs, while others are used for treating ADEs. Many of the highly-ranked terms appear only in a single or a handful of datasets; additional highly-ranked terms that appear in all 27 datasets – and conceivably important for detecting ADEs generally – include smårtá (Eng: pain), trött (Eng: tired), feber (Eng: fever) and utslag (Eng: rash). Named entity mentions of type ADE Cue were ranked somewhat lower (out of \( \sim 78k \)): reaktion (Eng: reaction) – 53, biverkan (Eng: side effect) – 332, läkemedelsbiverkan (Eng: drug reaction) – 855 and läkemedelsutlöst (Eng: drug-induced) – 19602. When inspecting
Table 2: Average accuracy (%) over 27 ADE datasets with different representations of clinical notes

| Representation                        | TF   | TF-IDF |
|--------------------------------------|------|--------|
| Unigrams                             | 83.05| 81.70  |
| Bigrams                              | 76.65| 75.98  |
| Trigrams                             | 68.13| 66.93  |
| Ngrams (Unigrams + Bigrams + Trigrams) | 79.47| 78.33  |
| Terms                                | 83.12| 81.47  |
| Unigrams + Terms                     | 83.09| 81.81  |
| Distributed (Unigrams)               | 81.13| 79.59  |
| Distributed (Terms)                  | 82.82| 82.99  |
| Distributed Ensemble (Unigrams)      | 82.23| 81.53  |
| Distributed Ensemble (Terms)         | 83.51| 82.82  |
| Distributed (Unigrams + Terms)       | 82.04| 80.63  |
| Distributed Ensemble (Unigrams + Terms) | 83.31| 82.30  |
| Unigrams + Terms + Distributed (Unigrams + Terms) | 83.89| 82.72  |

The average rank of terms of varying length, unigrams were ranked the highest, followed by bigrams, trigrams and ngrams with $n > 3$. Calculating the average rank of terms of various named entity types revealed that ADE Cue was ranked the highest, followed by Disorder, Body Part, Drug and Finding.

Table 3: Variable importance of terms

| Rank | Term (Swedish) | Term (English) | NE Type | Support |
|------|----------------|----------------|---------|---------|
| 1    | missbruk       | addiction      | Finding | 23      |
| 2    | bev-fl-iri     | bev-fl-iri     | Drug    | 2       |
| 3    | amfetamin      | amphetamine    | Drug    | 20      |
| 4    | cyanokit       | cyanokiti      | Drug    | 1       |
| 5    | l"akemedels-utlöst dystoni | drug-induced dystonia | Disorder | 1       |
| 6    | betapred       | betapred       | Drug    | 27      |
| 7    | intox          | intoxication   | Disorder | 22      |
| 8    | akut           | acute          | Disorder | 3       |
| 9    | dystoni        | dystonia       | Disorder | 27      |
| 10   | hepatit c      | hepatitis c    | Disorder | 27      |
| 11   | allergisk reaktion | allergic reaction | Disorder | 25      |
| 12   | tavegyl        | tavegyl        | Drug    | 25      |
| 13   | syrgas         | oxygen         | Drug    | 27      |
| 14   | amfetamin-missbruk | m"atbar sjukdom | Abuse   | 23      |
| 15   | stesosolid     | stesosolid     | Drug    | 26      |
| 16   | svullnad       | swelling       | Finding | 27      |
| 17   | kontrahera     | contract       | Finding | 1       |
| 18   | bltr vara      | blood pressure | Finding | 1       |
| 19   | stabil         | be stable      |         |         |
| 20   | klåda          | itching        | Finding | 27      |
|      | hjärtmuskel-inflammation | myocarditis | Disorder | 1       |

Table 4: The distribution of terms over classes

| NE Type | ADE | Not ADE |
|---------|-----|---------|
| Term / Episode | Tokens / Term | Term / Episode | Tokens / Term |
| Disorder   | 34.65 | 103.19  | 21.57  | 97.28  |
| Finding    | 124.14 | 28.80   | 68.07  | 30.83  |
| Drug       | 74.68  | 47.87   | 39.68  | 52.89  |
| Body Part  | 49.27  | 72.57   | 27.58  | 76.08  |
| ADE Cue    | 1.94   | 1839.01 | 0.86   | 2432.98|

4 Discussion

This study explored the use of 42 different representations of clinical notes from healthcare episodes for the automatic detection of adverse drug events. It was shown that combining local and global, distributed representations yielded the highest predictive performance. While the use of a simple unigram model worked well, performance quickly deteriorated as larger ngrams were used, most probably as a result of the ensuing sparsity. Interestingly, using only extracted terms outperformed the use of all unigrams, with the added benefit that the former is much lower-dimensional and thus computationally preferable. Even lower-dimensional – and denser – are the
distributed representations: in this case 200 with a single semantic space and 200 × 5 with the semantic space ensemble. A distinct advantage of distributed representations is their scalability, as the dimensionality does not grow with the size of the vocabulary, allowing more information to be exploited effectively, as demonstrated by the distributed ensemble of unigrams and terms. The best results were, however, obtained when combining local and ensembles of global, distributed representations. While the difference to using a simple unigrams model is not very large, it is interesting to note the bigger difference to using the commonly employed bag-of-words representation. The advantage of using a binary representation over TF or TF-IDF weighting was also somewhat surprising but can perhaps be attributed to the noisy nature of clinical text.

An advantage of using the terms representation is that, in comparison to the other representations – in particular the distributed ones – it lends itself to some degree of interpretability. While random forest belongs to a family of opaque models, inspection of variable importance provides some insight. It was not surprising that ADE Cue terms were, on average, ranked the highest, although somewhat more so that Body Part terms were ranked higher than Drug and Finding terms. When inspecting the distribution of terms over classes, however, it was confirmed that Drug and ADE Cue terms were common in ADE episodes than in non-ADE episodes, which seems intuitive. For future work, it would be interesting to study whether enriching the representation with factuality – including negation and uncertainty – and temporality would be lead to improved predictive performance.

Acknowledgments

This work was supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, volume 1, pages 238–247.

Leo Breiman. 2001. Random forests. Machine Learning, 45(1):5–32.

Emmanuel Chazard, Gregoire Ficheur, Stephanie Bernonville, Michel Luyckx, and Regis Beuscart. 2011. Data mining to generate adverse drug events detection rules. IEEE Transactions on Information Technology in Biomedicine, 15(6):823–830.

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. Journal of Biomedical Informatics, 42(2):390 – 405.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. In Proceedings of the Swedish Language Technology Conference (SLTC).

Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, Lars Juhl Jensen, and Søren Brunak. 2013. Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. Journal of the American Medical Informatics Association, 20(5):947–953.

Robert Eriksson, Thomas Werge, Lars Juhl Jensen, and Søren Brunak. 2014. Dose-specific adverse drug reaction identification in electronic patient records: temporal data mining in an inpatient psychiatric population. Drug safety, 37(4):237–247.

Salvador Garcia and Francisco Herrera. 2008. An Extension on” Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons. Journal of Machine Learning Research, 9(12).

Katja M Hakkarainen, Khadija Hedna, Max Petzold, and Staffan Hägg. 2012. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions–a meta-analysis. PloS one, 7(3):e33236.

Rave Harpaz, Krystl Haerian, Herbert S Chase, and Carol Friedman. 2010. Mining electronic health records for adverse drug effects using regression based methods. In Proceedings of the 1st ACM International Health Informatics Symposium, pages 100–107. ACM.

Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. Clinical Pharmacology & Therapeutics, 91(6):1010–1021.

Lorna Hazell and Saad AW Shakir. 2006. Under-reporting of adverse drug reactions. Drug Safety, 29(5):385–396.
Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravicius, and Martin Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5:6:1–25.

Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. 2015a. Identifying Adverse Drug Event Information in Clinical Notes with Distributional Semantic Representations of Context. *Journal of Biomedical Informatics*, in press.

Aron Henriksson, Jing Zhao, Henrik Boström, and Hercules Dalianis. 2015b. Modeling heterogeneous clinical sequence data in semantic space for adverse drug event detection. In *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE.

RL Howard, AJ Avery, S Slavenburg, S Royal, G Pipe, P Lucassen, and M Pirmohamed. 2007. Which drugs cause preventable admissions to hospital? a systematic review. *British Journal of Clinical Pharmacology*, 63(2):136–147.

Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigan H Shah. 2013. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov. 2015. *word2vec*: Tool for computing continuous distributed representations of words. https://code.google.com/p/word2vec/. Accessed: 2015-08-11.

Jonathan R Nebeker, Paul Barach, and Matthew H Samore. 2004. Clarifying adverse drug events: a clinician’s guide to terminology, documentation, and reporting. *Annals of internal medicine*, 140(10):795–801.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, 3:1–18.

Magnus Sahlgren and Rickard Cöster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 487. Association for Computational Linguistics.

Jürgen Staussberg and Joerg Hasford. 2011. Drug-related admissions and hospital-acquired adverse drug events in germany: a longitudinal analysis from 2003 to 2007 of icd-10-coded routine data. *BMC Health Services Research*, 11(1):134.

Marc Suling and Iris Pigeot. 2012. Signal detection and monitoring based on longitudinal healthcare data. *Pharmaceutics*, 4(4):607–640.