Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text

Eriksson, Robert; Jensen, Peter Bjødstrup; Pletscher-Frankild, Sune; Jensen, Lars Juhl; Brunak, Søren

Published in:
Journal of the American Medical Informatics Association

Link to article, DOI:
10.1136/amiajnl-2013-001708

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text

Robert Eriksson,1,2 Peter Bjøstrup Jensen,2 Sune Frankild,2 Lars Juhl Jensen,2 Søren Brunak1,2

ABSTRACT
Objective Drugs have tremendous potential to cure and relieve disease, but the risk of unintended effects is always present. Healthcare providers increasingly record data in electronic patient records (EPRs), in which we aim to identify possible adverse events (AEs) and, specifically, possible adverse drug events (ADEs).

Materials and methods Based on the undesirable effects section from the summary of product characteristics (SPC) of 7446 drugs, we have built a Danish ADE dictionary. Starting from this dictionary we have developed a pipeline for identifying possible ADEs in unstructured clinical narrative text. We use a named entity recognition (NER) tagger to identify dictionary matches in the text and post-coordination rules to construct ADE compound terms. Finally, we apply post-processing rules and filters to handle, for example, negations and sentences about subjects other than the patient. Moreover, this method allows synonyms to be identified and anatomical location descriptions can be merged to allow appropriate grouping of effects in the same location.

Results The method identified 1 970 731 (35 477 unique) possible ADEs in a large corpus of 6011 psychiatric hospital patient records. Validation was performed through manual inspection of possible ADEs, resulting in precision of 89% and recall of 75%.

Discussion The presented dictionary-building method could be used to construct other ADE dictionaries. The complication of compound words in Germanic languages was addressed. Additionally, the synonym and anatomical location collapse improve the method.

Conclusions The developed dictionary and method can be used to identify possible ADEs in Danish clinical narratives.

BACKGROUND AND SIGNIFICANCE
An unfortunate effect of any medical intervention is the risk of patients experiencing adverse events (AEs), which in the case of a drug intervention is known as an adverse drug event (ADE). The latter includes any adverse incident occurring during drug treatment, without necessarily implying a causal relationship with the treatment.1 These incidents are not limited to subjective descriptions by the patient and include quantitative changes in laboratory values.

It is estimated that ADE-related emergency department visits in the USA exceed 700 000 per year,2–3 and a UK study identified 6.5% of 18 820 hospital admissions as being directly or indirectly related to ADEs.4 These are conservative estimates for the total number of ADEs since less severe ADEs are often only recorded in the general practice, the patient may never seek medical attention or the effect is ignored. Apart from the human consequences of ADE-related morbidity and mortality, the substantial expenses for the healthcare system that result from these events4,5 provide a great incentive for a dedicated effort to record and reduce the number of ADEs.

Today the identification and quantification of ADEs starts during clinical drug trials and surveillance continues after market authorization. Post-approval safety monitoring is essential as many ADEs can go unnoticed through clinical trials, which never exactly reflect the clinical reality drugs are eventually used in. Spontaneous reporting is a usual approach to support the monitoring of adverse effects.6 This requires ADEs to be identified in the clinical setting, reported to a medical product agency or the pharmaceutical manufacturer and subsequently collected in databases such as the FDA Adverse Event Reporting System (FAERS) in the USA, EudraVigilance in the EU and the WHO VigiBase.6–7 Statistical patterns of drug–ADE co-occurrences are derived from such data and explored in order to refine drug safety profiles and detect previously unnoticed relationships. However, it is recognized that these databases are subject to biased reporting as well as gross under-reporting.8

Clinical narratives
Clinical text has some distinct properties that make information extraction particularly challenging and different from information extraction from scientific publications. Clinical narratives usually consist of short entries with diverse structures and styles, seldom conforming to standard grammar and containing more spelling and typing errors than published text.9 Abbreviations, shorthand and acronyms are common and their meaning is often ambiguous depending on the context. Another vital issue is the extensive use of negations to rule out clinical signs and references to subjects other than the actual patient.9–11

A number of text-mining tools have been developed to tackle the task of recognizing and extracting relevant clinical entities from clinical text.12–15 They typically rely on extensive dictionaries of concepts such as the UMLS or selected parts thereof, and implement various natural language processing (NLP) techniques.

Several studies have specifically examined automatic extraction of AEs from clinical texts. Murff et al16 used trigger words to screen discharge summaries for adverse medical events. Wang et al17 parsed discharge summaries using MedLEE and...
through filters identified drug-potential ADE relationships. Sohn et al\textsuperscript{18} extracted drug side effects from clinical text by using cTAKES. Honigman et al\textsuperscript{19} employed computer search methods to identify ADEs through multiple methods including text searching.

Critical for our work, the bulk of relevant research has focused on English text, largely owing to the availability of large lexical resources uniquely available in English, and the lack of a Danish ADE dictionary or terminology. Neither the WHO Adverse Reactions Terminology (WHO-ART),\textsuperscript{20} Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART)\textsuperscript{21} or Medical Dictionary for Regulatory Activities (MedDRA)\textsuperscript{22} have been translated to Danish. Although named entity recognition (NER) efforts for Danish and other Scandinavian languages do exist,\textsuperscript{23} these are not designed for clinical text. To our knowledge only one publication covers clinical text mining in Danish,\textsuperscript{24} where the authors used phenotypical descriptions to identify disease correlations from free-text patient records. In Swedish, another smaller language, NER has been employed to recognize disorders, findings and body structures by using a rule-and terminology-based system.\textsuperscript{25} The novelty in this work is that we use NER on Danish clinical narratives to identify possible ADEs. The shortage of tools and a dictionary prompted us to develop methods and solutions to overcome these problems.

**OBJECTIVE**

Improving and simplifying the steps involved in detecting possible ADEs as they occur in the clinical setting is important to improve ADE reporting and drug safety knowledge as well as the care of individual patients. In the clinical setting, care could be enhanced by reducing the risk of ADEs being overlooked and thus excluded in the clinical reasoning and decision-making process.

The automated detection and classification of possible ADEs in clinical narratives recorded in electronic patient record (EPR) systems can provide detailed, structured and time stamped data for use in the immediate clinical setting as well as in automated ADE reporting, EPR decision support systems or as the basis for further data-driven research into causal relationships between drugs and side effects.

It is important to emphasize that the WHO definition\textsuperscript{1} of an ADE does not imply an established causal relationship to a drug. Any adverse incident that temporally overlaps with a period of drug therapy is by definition an ADE. Only when causality is established is the term ‘adverse drug reaction’ (ADR) used. In this work we specifically focus on the detection of possible ADEs in clinical text; we seek to identify descriptions of events in the text that could possibly be associated with the use of a drug, but we do not attempt to establish if drug causality exists.

**MATERIALS AND METHODS**

A fundamental requirement for NER of possible ADE descriptions in clinical text is a dictionary of ADE terms. Since no such dictionary is available in Danish, we chose to create one rather than translate an existing resource. We estimated that this would be a comparable amount of work while giving us full flexibility to structure and optimize the dictionary.

We used the Danish summaries of product characteristics (SPCs) as the basis for our dictionary. These are the product summaries that manufacturers are legally required to supply in the EU and contain information about the product including its observed undesirable effects. The SPCs were chosen because drug manufacturers often translate information from a core data document maintained in one of the languages covered by the terminologies. This means the same term has been translated several times by the manufacturers and therefore there are alternative and non-standardized translations, including layman terms, thus expanding the coverage of the dictionary.

**Danish ADE dictionary creation**

The undesirable effects section of 7446 Danish SPCs for products with national or centralised authorization for marketing in Denmark was chosen as the initial source of ADE descriptions for the dictionary. A sentence splitter was used to identify all unique lowercase normalized sentences. This reduced the manual task of extracting all ADE descriptions from the set of unique sentences. All ADE descriptions were extracted, including those in Danish, medical Latin and some in English. Descriptions in English were not added if they had similar spelling as descriptions from the other two languages but with different meaning. Duplicate descriptions were removed and the remaining unique descriptions were manually validated, preserving all spelling variants. This resulted in 21 342 unique ADE descriptions ranging from single word ADEs like *galactorrhea* to multi-word ADEs like *rash in the palm of hand* (see figure 1A). As the dictionary is aimed at noxious and unintended effects, we only included ADEs and not beneficial side effects.\textsuperscript{1}

**Dictionary groups and blacklist**

We created two versions of the ADE dictionary. A baseline version where every extracted unique SPC description formed a separate lexeme and a more sophisticated group-based version designed to allow for a higher degree of flexibility in ADE detection. Details of the group-based dictionary creation are described below and illustrated in figure 1A,B.

All ADE descriptions extracted from the SPCs where manually assigned to the seven groups shown in figure 1B to enable the identification of additional ADE descriptions through post-coordination of different groups as described later in this section. In many cases this group assignment splits multi-word descriptions into individual lexemes that were grouped individually. For example *rash in the palm of hand* was split into the groups: location—*palm of hand*, event—*rash* and preposition—*in*. Other descriptions are either single words or represent ADEs that cannot sensibly be split, like *galactorrhea* and *grand mal*. Furthermore, each group was augmented with additional words not seen in the SPCs, such as additional prepositions and more anatomical locations from the BRENDA Tissue Ontology.\textsuperscript{27} Finally, a number of spelling and inflection variants were introduced using simple rules to improve detection. The seven groups are described in detail below and examples are included in figure 1.

Parallel to the manual assignment of SPC descriptions to groups, the resulting lexemes were assigned identifiers. Within group synonyms, spelling variants and inflectional variants of the same ADE concept were assigned a common identifier.

- **Independent event**. These lexemes describe a particular event that by itself is a possible ADE. This category can consist of a single word or more in a specific order. Lexemes with either of the prefixes *bypo* or *byper* were included in the dictionary with both prefixes.
- **Abbreviation**.Abbreviations are treated in a similar fashion as the independent event group, but are grouped separately to allow for multiple meanings of ambiguous abbreviations.
- **Localized event**. Lexemes in the localized event group only qualify as a possible ADE during post-coordination if they are tagged together with a location. Many lexemes such as *itching*
are grouped both as localized event and as independent event to reflect their ability to specify an ADE alone or in combination with a location.

- **Location.** Members of the group describe anatomical locations and were mapped to the BRENDA Tissue Ontology to enable ADE aggregation on different levels of anatomical detail as described later. The group also contains lexemes such as mental that are not strictly anatomical locations.

- **Laboratory event.** Lexemes indicating change in a non-numerical result of a laboratory value are assigned to this group. Additionally, versions of positive, like pos and +, are included.

- **Laboratory value.** Full length and abbreviated, clinical chemistry, hematology and clinical microbiology laboratory variables are included. Sample origin information for blood, plasma, serum, cerebrospinal fluid, urine or location not specified is likewise included in full length and abbreviated form. All laboratory variables and sample origin are combined through combinatorial expansion to all possible expressions. Specific laboratory values are marked and later allowed to be combined with the laboratory event positive; this is described in more detail later.

- **Preposition.** The preposition group contains all relevant prepositions which are used in post-coordination to link location with event or laboratory value with laboratory event. Positions describing within location placement, like right and lateral, are added to the prepositions through combinatorial expansion, but we exclude combinations indicating relative positions outside the structure, like distal to. Likewise, positions like adjacent to are not added as they reflect relative positions that often are not within the same anatomical structure.

In addition to these seven groups the dictionary contains four groups (see figure 1E) which are used for filtering out possible ADEs.

- **Negation.** This group consists of negations and words giving rise to uncertainty about the possible ADEs in the sentence. Some lexemes in this group are not negations in the traditional sense of the word but rather reflect the common use of uncertain and speculative descriptions in the clinical narratives.

- **Other subject.** Lexemes indicating the information in the sentence could be about subjects other than the patient.

- **Undesirable effects information.** It is common and mandatory for a physician to explain the possible undesirable effects of any new drug treatment to the patient and document this in the record before treatment begins. This typically has a standardized content with a few strongly indicating words such as patient informed about side effects that we assign to this group. This group allows us to disqualify possible ADEs in a sentence where a lexeme from this group is also present.

- **Temporal trigger.** This group contains a large number of combinations of words and numbers indicating a past time point. We use this trigger group to exclude ADEs that are likely past events.
A corpus specific blacklist was developed in parallel to the
tagger development and testing phase where false positive con-
cepts were identified. The developed blacklist is small in com-
parison to the dictionary and contains only 59 unique concepts
such as the Kage (Danish city) identified as potassium-increase
and smore (smear) identified as pain-ear. All entries were
derived from words incorrectly interpreted by the method as
different meaning compound words, laboratory value abbrevia-
tions or Danish geographical locations.

Corpus tagging
A modified version of the Reflect tagger28 was used to tag the
entire corpus based on the dictionary and blacklist. The original
tagger was developed for fast tagging of English language texts,
and required a few modifications to enable efficient tagging of
Danish clinical text. Most importantly, the tagger was set up to
tokenize on every character and allowed tags across word
boundaries but not sentence boundaries. Sentences were split
using the Reflect sentence splitter. In each sentence we used a
tagging window of a maximum of 50 characters. Single charac-
ter tokenization was used to identify lexemes in compound
words. Likewise each lexeme in falsely separated compounds
was tagged individually. This allowed us to merge compounds
and falsely separated compounds in the later post-coordination.
The full compound words do not need to be in the dictionary
as long as the individual lexemes are. Both compounded and
separated lexemes were combined in the following post-
coordination to form coordinated terms. Falsely compounded
words were identified providing a version was present in the
dictionary, because the dictionary look-up ignores spaces and
hyphens.

Artifact filtering
Although beneficial for handling compound words, the tagger
set-up produces a large number of artifacts caused by matches
to subparts of words and matches that inappropriately span
word boundaries. The first corrections were made in the tagging
engine, which ignores matches across spaces where the begin-
ning or ending consists of a single character or a number of
short common words separated by space. The remaining arti-
facts were removed by specific filtering rules for individual dic-
tionary groups, described in detail later. All dictionary matched
terms need to have a word boundary before and after the
matched lexeme, except for post-coordination groups where
only a boundary before or after the match is required.

Post-coordination and synonymous ways of writing ADEs
Post-coordination is the process of detecting possible ADEs that
are not in the dictionary by combining multiple tagged lexemes
in close proximity according to certain rules. Post-coordination
combined lexemes of the group localized event with lexemes
from group localized event, and laboratory value lexemes were
combined with laboratory event lexemes. Lexemes from the
preposition group were allowed between tags. Identified diction-
ary groups not having a corresponding dictionary group next to
them or lacking word boundaries in the appropriate places were
discarded.

Post-coordination also allows one localized event or labora-
tory event to point at multiple locations or laboratory values,
separated by a comma or the conjunction and. For example, the
phrase oral and gastric bleeding results in the pairing of the two
locations with the single localized event, creating the pairs oral-
bleeding and gastric-bleeding.

Post-coordination assures the equality of synonymous ways to
write the same ADE. This was done by only retaining informa-
tion about localized event and laboratory event combined with
location and laboratory value, illustrated in figure 2. Any order of
coordinated tags and the presence or absence of prepositions
were treated as identical. The laboratory event positive can be
combined with specific laboratory values to create specific clinically
relevant pairs, like positive anti-nuclear antibodies.

Negations and negatives filtering
The next post-processing step is the removal of tagged lexemes
disqualified by any of the negative filtering dictionary groups
shown in figure 1F. We used a within sentence negation scope of
a maximum of six words, disqualifying any possible ADE that
succeeded negations or other subjects. If the conjunction but
appears between the disqualifying lexeme and the possible ADE,
the negation scope is reduced to the conjunction. Although Danish
grammar does allow negations to succeed possible ADEs, manual inspection showed that this was hardly ever the
case in the hospital clinical texts. Possible ADEs identified any-
where in a sentence containing undesirable effects information
or a temporal trigger were also disqualified.

Location collapse and synonyms
ADE locations were collapsed using the BREnda Tissue
Ontology27 structure, which maps different levels of anatomical
detail and enables aggregation into organs and organ systems as
shown in figure 3. A separate input file, which can be modified,
controls the details of this collapsing. Alternative spellings, syno-
yms and inflectional variants were aggregated to the same iden-
tifier. This means that location variants like abdominal and
abdomen are assigned to the same identifier, likewise decrease
and reduce from the laboratory value group share one identifier.
These post-coordination conversions enabled us to aggregate dif-
ferent variations of coordinated terms as illustrated in figure 4.

EPR test corpus
The performance of the method was assessed on a corpus con-
sisting of all clinical narratives in an EPR system acquired from a
collaborating Danish psychiatric hospital. The corpus contains
notes on 6011 patients collected over 12 years and consists of
approximately 250 million words.

There are multiple character encodings in the corpus, most
likely due to different methods of entering text by the users,
such as transferring text from external sources.9 We therefore
standardized both the dictionary and the corpus to single-byte
ASCII characters before processing. Special Danish characters,
which are not covered by the encoding, were transformed to
ASCII character strings along with any other non-ASCII charac-
ters. By transforming both the dictionary and the corpus, the
risk of identifying false tags was limited and we did not observe
any errors arising from ASCII conversion.

RESULTS
The complete test corpus was mined using both the baseline dic-
tionary consisting only of the original ADE descriptions
extracted from the SPCs and the group-based dictionary. This
allowed us to evaluate the added value of the more advanced
approach.

Validation
To evaluate the performance of our pipeline, we created a valid-
atation text corpus consisting of 200 randomly selected patient
notes representing 181 different patients. Two annotators (the
first and second authors), one clinical pharmacist and one health
informatician, independently annotated the validation records
according to our stated goal of identifying possible ADE men-
tions without requiring a causal drug relationship. Thus all men-
tions of symptoms, signs or clinical terms with a direct,
non-negated link to the patient that could under some circum-
stance constitute an ADE were annotated as a possible ADE in
the test corpus by the annotators. A total of 216 ADEs were
annotated in this process with an inter-annotator agreement of
0.76 (Cohen’s κ). Comparison of the identified possible ADEs
of the text mining pipeline to the consensus reached by the
annotators resulted in precision and recall of 89% and 75%,
respectively, for the group dictionary and 90% and 47%,
respectively, for the baseline dictionary.

Half of the false positive matches produced by the group
dictionary text mining were explained by descriptions of treatments
and addresses tagged by the pipeline (eg, ‘foot center/diabetes
team’ where diabetes was tagged) and not per se events the patient
experienced. Missed descriptions of other subjects and negations
comprised 20% of false positives (eg, ‘He also has depression’
where in other sentences it is clear he is not the patient). The
remaining 30% of the false positives were equally divided between
a lack of disqualification of matches found in adverse effect inform-
gation given to the patient prior to treatment (eg, edema in ‘Most
common side effects are edema, sedation and nausea’) and misin-
terpretations (eg, the Danish words for withdraw in ‘withdraw
money’ and swollen in ‘swollen knee’ are identical).

False negatives from the text mining were 91% explained by
additional ways of describing a possible ADE not included in
the dictionary. No cases of errors due to misspellings, alternative
spellings or inflections were identified. The remaining 9% of
false negatives was explained by the use of the word for, which
is used to disqualify possible ADEs as indications for treatment.
We knowingly included this broad word, not only used to
describe indications, to produce a conservative output.

**Recognized and identified concepts in the full test corpus**

Text mining the entire corpus using the developed group dic-
tionary resulted in a total of 85 049 343 matched lexemes.
Artifact filtering, negative group filtering and post-coordination
resulted in 1 970 731 recognized possible ADE concepts rep-resented by 35 477 unique concepts and 11 641 concepts after
synonym reduction. Using the baseline dictionary resulted in
1 729 746 recognized concepts, which dropped to 1 437 849
possible ADE concepts after negative filtering. Thus the group
dictionary improves ADE recognition by 37% while maintaining
a precision of 89% as indicated by the validation above.

The 10 most recognized concepts using the baseline diction-
ary and the group dictionary are presented in table 1. Only the
group dictionary can identify synonymous concepts since the
full SPC descriptions of the baseline dictionary were not
assigned descriptive identifiers or post-coordinated.

Table 2 shows the contribution of different dictionary groups
to the detection of possible ADEs. The independent event
group was responsible for 74% of all identified concepts. This is
not unexpected as many possible ADEs are single word descrip-
tions or represent ADEs that cannot sensibly be split.

**Location collapse**

Without the location collapse feature, 242 identifiers represent-
ing 1213 ways to write locations in the corpus were used in
post-coordination. This was reduced by 10% to 217 identifiers
when the current level-of-detail location collapse was
implemented.

**False positive reduction**

The majority of falsely tagged lexemes were removed by artifact
filtering and during post-coordination. These two steps reduced
the 85 049 343 matches by a massive 96.8%, leaving 2 721 578
matches. Another 28% were disquali-
ed concepts. This is
enlist the equality of two different ways of writing the same ADE and
synonymous coordinated terms are merged into one common term. Order and possible prepositions used during post-coordination were excluded.

Figure 2  **Synonymous coordinated terms.** Lexemes from the groups **localized event and laboratory event** are combined with **location** and **laboratory values**, respectively, to produce coordinated terms. The method identifies the equality of two different ways of writing the same ADE and synonymous coordinated terms are merged into one common term. Order and possible prepositions used during post-coordination were excluded.

Figure 3  **Location collapse.** Locations are collapsed and merged into a single identifier.
As demonstrated, there was a significant gain in performance when a group-based dictionary was used in combination with post-coordination and compound handling over a baseline dictionary consisting of original full-length SPC descriptions. An additional 37% possible ADEs were identified by the group dictionary, corresponding to an increase in recall from 47% to 73% according to the validation set. Importantly, precision remained unchanged at about 90% since both dictionaries were sensitive to the same types of false positives.

Although the described methods were constructed for a Danish context, the high similarity among Scandinavian languages means that it would be possible to transfer the analysis to the other languages. Only modifications to the dictionary would be needed, which in large part could be recycled since the Scandinavian languages share many words and a similar syntax. We also believe that the suggested method provides a simple, yet efficient, system for detecting ADEs in languages where no ADE dictionary is available. We have no reason to believe that our pipeline would perform significantly differently on a different corpus, since the dictionary is based on the SPCs of all authorized drugs with no special targeting of the psychiatric domain.

In addition, we provide an efficient way to detect compound lexemes by using single character tokenization combined with dictionary look-up that ignores spaces and hyphens. This approach should allow this process and method to be converted to other compound-rich continental Germanic languages. It does produce a high number of artifact lexeme matches, but these are easily filtered out by a relatively small set of hand-crafted filters.

A unique feature of the method is the possibility of discovering and aggregating equivalent post-coordinated ADE descriptions by using the relationships of the constituent lexemes, for example, *renal failure* equals *failure of kidney*. Another is the aggregation of relevant anatomical structures by using the anatomical hierarchy of the BRENDA Tissue Ontology, for example, the previously described multiple tooth identifiers (figure 3) collapsed to one. Location collapse reduced the identifiers by 10%, providing the valuable ability to merge locations and hence better calculate the effects in a particular location. The use of synonym reduction and location collapse reduced 26,774 uniquely matched text strings to 9,484 identifiers only occupying approximately one third of the original identifier space.

It is important to state that the objective of the method presented here is to identify possible ADE descriptions in clinical text. Determining whether a possible ADE is in fact associated with a drug, or is rather a description of a symptom, diagnosis or indication, is not currently pursued and is a limitation of the method. That task would require temporal knowledge of administered drugs to be integrated probabilistically with the detected mentions of possible ADEs. Such information could also be obtained from text mining or perhaps preferably from structured prescription data. A further limitation is that close to one quarter of all manual possible ADE annotations in the validation set are not found by the method, which is largely due to the absence of the lexeme from the dictionary. Despite the simplicity of the rigid rule-based approach, more advanced filters would likely be beneficial, such as a more flexible negation

---

**Table 1** The 10 most recognized concepts in the corpus

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Possible ADE</th>
<th>Recognized/identified concepts</th>
<th>Unique ways identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group dictionary</td>
<td>Anxiety</td>
<td>128839</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Sedation</td>
<td>99191</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Pain</td>
<td>89660</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>Anger</td>
<td>75623</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Unrest</td>
<td>69322</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Auditory</td>
<td>65888</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>hallucination</td>
<td>59435</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Psychosis</td>
<td>41040</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Paranoia</td>
<td>36302</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Irritation</td>
<td>33673</td>
<td>20</td>
</tr>
<tr>
<td>Baseline dictionary</td>
<td>Anxiety</td>
<td>108049</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Psychosis</td>
<td>47725</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Unrest</td>
<td>34848</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Pain</td>
<td>32291</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Paranoia</td>
<td>29358</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Suicidal thoughts</td>
<td>23856</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Adverse effect</td>
<td>19589</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Headache</td>
<td>19487</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Restless</td>
<td>19868</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Schizophrenia</td>
<td>17962</td>
<td>–</td>
</tr>
</tbody>
</table>

For the group dictionary, the table shows the number of unique ways each possible ADE was identified. ADE, adverse drug event.

---

**Table 2** Dictionary group contributions to the total of 1970731 recognized concepts

<table>
<thead>
<tr>
<th>Dictionary group</th>
<th>Identified concepts</th>
<th>Unique text strings</th>
<th>Unique identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent event</td>
<td>1449924</td>
<td>5108</td>
<td>1707</td>
</tr>
<tr>
<td>Localized event and location</td>
<td>464052</td>
<td>26774</td>
<td>9484</td>
</tr>
<tr>
<td>Laboratory event and laboratory value</td>
<td>48543</td>
<td>3783</td>
<td>452</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>8212</td>
<td>37</td>
<td>33</td>
</tr>
</tbody>
</table>
implementing the method. The presented way of building a group-based dictionary and ensure it can be generalized.

Additionally, the presented method has only been verified on a psychiatric corpus and needs to be tested on other domains to ensure it can be generalized.

Dictionary construction and method development required close to 1 man-year, with each part taking about half a year. The presented way of building a group-based dictionary and implementing the method’s rule-based filters should allow others to recreate this approach in significantly less time.

CONCLUSION
We believe the dictionary and method presented here provide a solid foundation for more advanced text mining-based analyses in the future, as an example to associate undesirable effects and drugs. Based on our results, we believe that EPRs represent a valuable source of information on undesirable effects that should be exploited more fully. Our system provides a text-mining method for languages where an ADE dictionary and tools are not available, to allow the further exploration of ADEs documented by health professional in EPRs.

Acknowledgements The authors would like to thank Heiko Horn and Anders Jensen for valuable technical assistance.

Contributors RE, PBJ, LIJ and SB designed the study. RE designed the dictionaries, post-coordination, filters and location collapse. RE, SF and LIJ modified the reflect tagger and standardized characters. RE and PBJ analyzed the data and wrote the manuscript. All authors critically reviewed and approved the manuscript.

Funding The Danish Council for Strategic Research, the Novo Nordisk Foundation and the Villum Foundation funded this study.

Competing interests None.

Ethics approval The Danish National Board of Health approved this study.

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

Table 3 Negative filtering disqualification and dictionary group contribution

<table>
<thead>
<tr>
<th>Dictionary group</th>
<th>Contribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>56.5</td>
</tr>
<tr>
<td>Undesirable effects information</td>
<td>39.3</td>
</tr>
<tr>
<td>Other subject</td>
<td>3.9</td>
</tr>
<tr>
<td>Temporal trigger</td>
<td>0.3</td>
</tr>
</tbody>
</table>

REFERENCES
Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text

Robert Eriksson, Peter Bjødstrup Jensen, Sune Frankild, et al.

J Am Med Inform Assoc 2013 20: 947-953 originally published online May 23, 2013
doi: 10.1136/amiajnl-2013-001708

Updated information and services can be found at:
http://jamia.bmj.com/content/20/5/947.full.html

These include:

References
This article cites 20 articles, 8 of which can be accessed free at:
http://jamia.bmj.com/content/20/5/947.full.html#ref-list-1

Open Access
This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
Open access (111 articles)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/