BJCP British Journal of Clinical Pharmacology

DOI:10.1111/j.1365-2125.2011.04153.x

# Using text-mining techniques in electronic patient records to identify ADRs from medicine use

**Pernille Warrer,[1,2,3] Ebba Holme Hansen,[1,2,3] Lars Juhl-Jensen[4] & Lise Aagaard[1,2,3]**

*[1]Department of Pharmacology and Pharmacotherapy, Section for Social Pharmacy, Faculty of Pharmaceutical Sciences, University of Copenhagen, [2]FKL-Research Centre for Quality in Medicine Use, [3]Danish Pharmacovigilance Research Project (DANPREP), and [4]The Novo Nordisk Foundation Center for Protein Research (CPR), Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark*

**Correspondence**
Ms Pernille Warrer, Department of Pharmacology and Pharmacotherapy, Section for Social Pharmacy, Faculty of Pharmaceutical Sciences, University of Copenhagen Universitetsparken 2, DK-2100 Copenhagen, Denmark.
Tel.: +45 3533 6438
Fax: +45 3530 6050
E-mail: pwa@farma.ku.dk

**Authors' contributions**
PWA, EHH and LA designed the study; PWA conducted the literature search and extracted data; PWA, EHH, LJJ and LA analyzed the data. All authors read and approved the final version of the manuscript.

**Keywords**
adverse drug reactions, electronic patient records, natural language processing, pharmacovigilance, text mining

**Received**
23 August 2011
**Accepted**
22 November 2011
**Accepted Article Published Online**
29 November 2011

This literature review included studies that use text-mining techniques in narrative documents stored in electronic patient records (EPRs) to investigate ADRs. We searched PubMed, Embase, Web of Science and International Pharmaceutical Abstracts without restrictions from origin until July 2011. We included empirically based studies on text mining of electronic patient records (EPRs) that focused on detecting ADRs, excluding those that investigated adverse events not related to medicine use. We extracted information on study populations, EPR data sources, frequencies and types of the identified ADRs, medicines associated with ADRs, text-mining algorithms used and their performance. Seven studies, all from the United States, were eligible for inclusion in the review. Studies were published from 2001, the majority between 2009 and 2010. Text-mining techniques varied over time from simple free text searching of outpatient visit notes and inpatient discharge summaries to more advanced techniques involving natural language processing (NLP) of inpatient discharge summaries. Performance appeared to increase with the use of NLP, although many ADRs were still missed. Due to differences in study design and populations, various types of ADRs were identified and thus we could not make comparisons across studies. The review underscores the feasibility and potential of text mining to investigate narrative documents in EPRs for ADRs. However, more empirical studies are needed to evaluate whether text mining of EPRs can be used systematically to collect new information about ADRs.

## Introduction

Adverse drug reactions (ADRs) represent a major public health problem estimated to account for up to 6.5% of all hospital admissions [1], 28% of all visits to casualty departments [2] and 6.4% of hospital fatalities [3]. Many rare and long term ADRs are not known at the time of marketing due to the limitations in safety exploration during preclinical and clinical research. Randomized clinical trials have limitations including small sample size, short duration and restrictive inclusion criteria. Furthermore, randomized clinical trials represent hypothesis testing of efficacy and not safety [4, 5]. Spontaneous reporting of ADRs by healthcare professionals is the most important approach in detecting new, rare and long term ADRs. However, not all ADRs are actually detected this way, as shown by the appearance of unexpected and serious ADRs many years after marketing [6, 7]. Hence, there is a need for alternative approaches to identifying ADRs not known at the time of marketing. Electronic patient records (EPRs) along with the development of powerful computers and new statistical algorithms hold great potential for detecting unknown ADRs. Data mining has been explored in recent years and has been described as a successful approach for extracting valuable information from EPRs on ADRs [8–10]. However, focus has been on coded and structured data and data

available in narrative EPR documents have been left out. Text mining, referred to as 'the process of deriving high quality information from text' [11] has been successfully employed in processing clinical notes within various medical specialties (e.g. radiology [12], asthma management [13], smoking cessation [14] and psychiatry [15]). Very few studies, however, have explored the use of narrative EPR data to investigate ADRs.

To the best of our knowledge, no previous literature review focused explicitly on the use of text-mining techniques in EPRs to identify ADRs from medicine use.

The aim of this article is to review studies using text-mining techniques to investigate narrative documents stored in EPRs for ADRs by posing the following questions: Which text-mining techniques were used in EPRs to investigate ADRs? How did they perform in terms of identifying ADRs? What were the frequencies and types of ADRs identified in the included studies?

## Methods

### *Literature search*

Literature searches were performed in PubMed, Embase, Web of Science and International Pharmaceutical Abstracts from origin and updated until July 2011. The complete databases were searched without restrictions for the combinations of search terms shown in Table 1 (see Appendix 1 for further details). The reference lists of the identified articles were manually checked for additional potentially relevant articles.

### *Study selection*

One author (PWA) retrieved and assessed articles identified during the literature searches based on the information provided in the title, abstract and descriptor/MeSH terms. To be considered relevant for this review, articles had to contain empirical data about text-mining EPRs to identify ADRs. We excluded articles exploring data mining exclusively, assessing text-mining systems for purposes other than ADR detection or using data sources other than EPRs. Articles exploring EPR data in combination with other data were included if sufficient data were provided to allow evaluation of the EPR data. Also excluded were articles focusing on medical error, ADR prevention or drug prescription strategies, or studies evaluating the performance of diagnostic tests and treatment procedures.

### *Data extraction*

Information was extracted from all full text articles selected for further assessment. Parameters extracted included study populations, settings, EPR data sources, frequencies and types of ADRs identified, medicines associated with ADRs, text-mining algorithms used and their performance. One author (LA) checked the extracted data in included articles, while another author (EHH) checked excluded articles. In the event of disagreement, consensus was reached through discussion.

We used the EU definition of an ADR: 'A response to a medicinal product which is noxious and unintended and which occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease' [16].

## Results

The literature searches identified 205 potentially relevant articles. A total of thirty-six articles were retrieved for further assessment and data extraction. Thirty-two of these

**Table 1**

Search terms

| PubMed | AND | Embase | AND | Web of science | AND | IPA |
|---|---|---|---|---|---|---|
| | | | ↓ | | | |
| **Data mining** | **AND** | | Adverse drug event | | **AND** | Clinical record |
| | | | | | | Computerized medical record |
| **Information extraction** | | | Adverse drug reaction | | | Computerized patient record |
| | | | | | | Electronic health record |
| **Knowledge acquisition** | | | Adverse event | | | Electronic medical record |
| | | | | | | Electronic patient record |
| **Knowledge extraction** | | | Adverse reaction | | | Free-text medical record |
| | | | | | | Medical record |
| **Mining** | | | Drug toxicity | | | Medical records system |
| | | | | | | Narrative patient record |
| **Text mining** | | | Pharmacovigilance | | | Patient record system |
| | | | | | | Unstructured textual information |

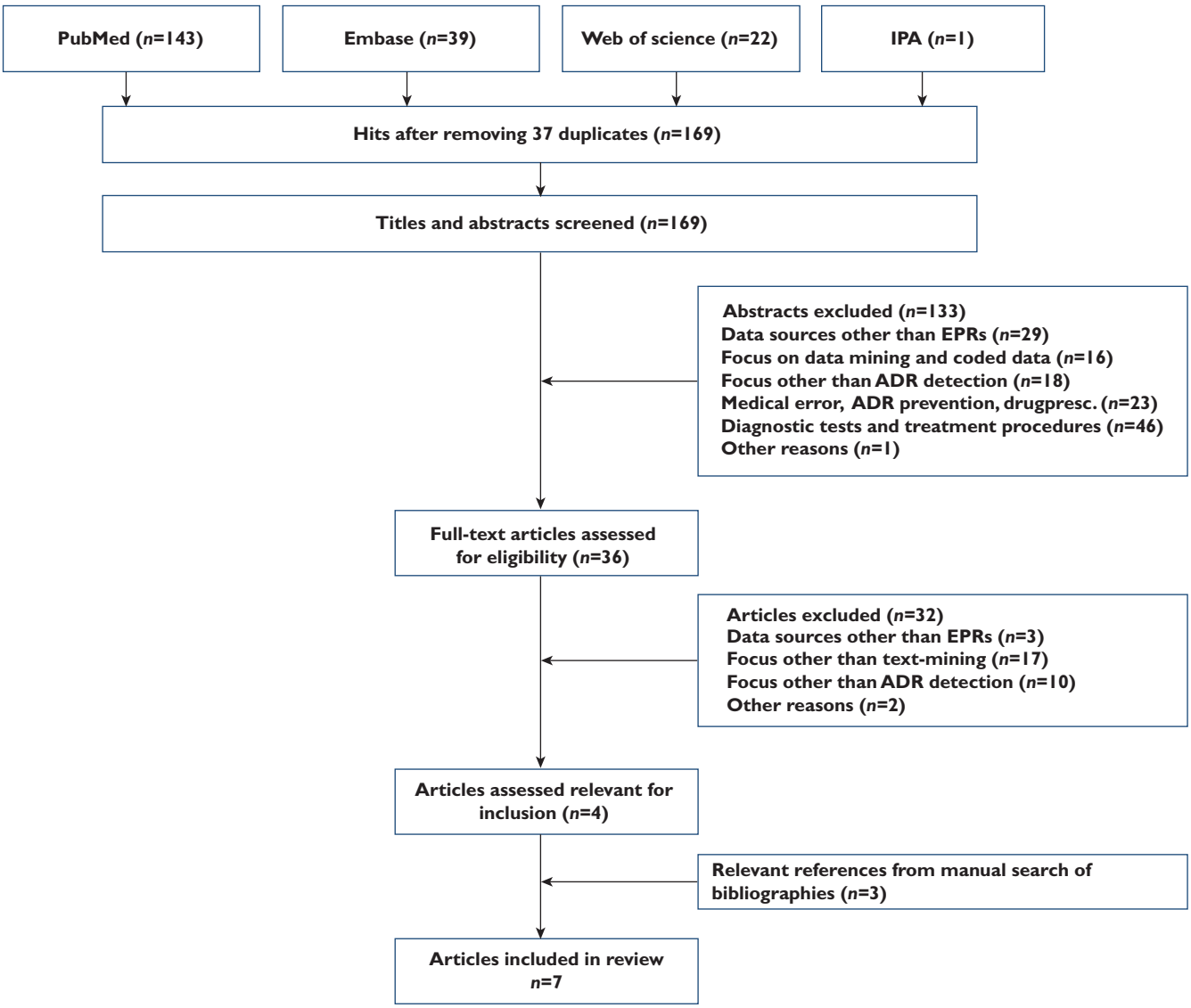


**Figure 1**
Flow chart of selection and review process for included/excluded articles

articles were excluded, mainly because they did not focus on identifying ADRs ($n = 10$) or because they described techniques for data extraction other than text mining ($n = 17$). Five articles were excluded for other reasons. A flow chart of the selection and review process, including reasons for exclusion, is shown in Figure 1. There were no cases of disagreement and, consequently, four articles were selected for inclusion in the review. Manual search of reference lists resulted in an additional three articles, bringing the total number of included articles to seven.

## *Study characteristics*

A summary of study characteristics is displayed in Table 2. Articles were published from 2001 onwards, the majority between 2009 and 2010. However, the data on which the studies were based were collected earlier. All of the articles were based on American study populations and derived mainly from two hospitals, the Brigham and Women's Hospital in Boston, Massachusetts [17, 18] and the New York Presbyterian Hospital [10, 19, 20]. For the remaining two studies, one included patients from a large multi-specialty group practice [21], while one did not state setting [22]. Four articles used inpatient data [10, 18–20] and three used outpatient data [17, 21, 22]. Study populations were broad and generally not defined in terms of age and gender, with the exception of one article that studied outpatients aged 65 years or older [21]. The number of patients included ranged from a total of 424 [18] to 381 084 [21]. However,

## Table 2

Summary of study characteristics

| References | Design | Setting | Health sector | Patient population | Data collection | Study period (months) | EPR system* | EPR data source(s) |
|---|---|---|---|---|---|---|---|---|
| **Honigman *et al.* [17]** | Cohort study | Brigham and Women's Hospital, USA | Outpatients | All patients with a primary care physician coming for care ($n$ = 15 665) | 1995–1996 | 12 | Brigham Integrated Computer System | Patient demographics, medical problem lists, ICD-9 claims, patient allergies, medication history and all clinic visit notes |
| **Murff *et al.* [18]** | Cohort study | Brigham and Women's Hospital, USA | Inpatients | Patients admitted to general medicine and medical subspecialty services ($n$ = 424) The study population was divided into cohort 1 ($n$ = 424) and cohort 2 ($n$ = 2826). Results provided, however, were mainly based on cohort 1. | January–June 2000 | 6 | Brigham Integrated Computer System | Discharge summaries |
| **Field *et al.* [21]** | Cohort study | Large multispecialty group practice affiliated with a New England-based health maintenance organisation, USA | Outpatients | Patients aged 65 years or older receiving health care services delivered by the group practice in the ambulatory setting ($n$ = 31 757 per month) | July 1999–June 2000 | 12 | Not stated | Discharge summaries, casualty department notes, computer-generated signals, electronic clinic notes and incident reports |
| **Hazlehurst *et al.* [22]** | Cohort study | Not specified, USA | Outpatients (and casualty department visits) | Outpatients and patients visiting the casualty department who had any immunization recorded ($n$ = not stated) | January–April 2004 | 4 | EPR of a large health maintenance organization | Clinicians' text notes |
| **Wang *et al.* [10]** | Cohort study | New York Presbyterian Hospital, USA | Inpatients | Patients admitted (not further specified) ($n$ = not stated) | 2004 | 12 | Clinical information system at New York Presbyterian Hospital | Discharge summaries ($n$ = 25.074) |
| **Haerian *et al.* [19]** | Cohort study | New York Presbyterian Hospital, USA | Inpatients | All patients with a serum creatine kinase >5 times normal ($n$ = 687) | 2004–2008 | 48 | Not stated | Discharge summaries |
| **Wang *et al.* [20] **** | Cohort study | New York Presbyterian Hospital, USA | Inpatients | Patients admitted (not further specified) ($n$ = not stated) | 2004 | 12 | Clinical information system at New York Presbyterian Hospital | Discharge summaries ($n$ = 25.074) |

*Electronic patient records system.

**Uses the same data as in Wang *et al.* [10] but with a different focus and methodology.

the number of patients included was only clearly stated in four of the seven included articles. Data collection periods ranged from 4 months to 4 years. Data sources included various textual EPR documents with an overweight of discharge summaries, which four studies used [10, 18–20]. One study used clinicians' text notes [22] while the remaining two studies used combinations of various textual EPR data [17, 21].

## *Text-mining techniques and performance*

Table 3 summarizes the text-mining techniques applied to EPRs to detect ADRs as well as their performance. Honigman *et al*. combined four different computer search methods (see Table 3) to determine the frequency and types of ADRs in outpatients. The sensitivity and specificity of the combined methods were estimated to be 58% (95% CI 18, 98) and 88% (95% CI 87, 88), respectively. The positive predictive value (PPV) was 7.5% (95% CI 6.5, 8.5) and the negative predictive value (NPV) was 99.2% (95% 95.5, 99.98). Of the four search methods, free text searching of clinic notes accounted for 90.6% of the ADRs detected, although the PPV was relatively low at 7.2% (95% CI 6.8, 7.5) [17]. Field *et al*. used similar methods in their study of ADRs among older persons in an ambulatory setting. The study used signals from multiple sources (see Table 3) and found that free text searching of clinic notes identified the highest percentage of ADRs (39%) with a PPV of 12% [21]. Murff *et al*. developed a computerized screening tool using simple keyword queries to search discharge summaries for trigger words representing a broad range of events including ADRs in inpatients (divided into cohort 1 and cohort 2). Events in the study were defined as injuries resulting from medical management. The most common events detected (52%) were ADRs. However, the reported sensitivity of 69%, specificity of 48%, and PPV of 52% were based on the detection of all events found in the study and were not reported for ADRs alone. When the electronic screening tool was coupled with manual discharge summary review, sensitivity was 64% (95% CI 61, 67), specificity 85% (95% CI 80, 90) and PPV 78% (95% CI 71, 85). These figures are based on results from cohort 1 only (see Table 3) and were not available for cohort 2 [18].

Hazlehurst *et al*. developed an automated classification system, MediClass, using NLP and knowledge-based techniques to identify possible vaccine-related ADRs, in particular those of gastrointestinal origin, in clinical notes of outpatient visits recorded by clinicians. Compared with standard methods that use diagnosis and utilization codes, the NLP-based system was estimated to identify approximately four times as many vaccine-related ADRs with a PPV of 74%. PPV for the code-based method was 31% [22]. MedLEE, a similar NLP-based system, was used by three other studies [10, 19, 20]. Wang *et al*. used MedLEE on a collection of discharge summaries to identify potential ADRs in inpatients. Seven drugs (ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone, ACE inhibitors) were selected to evaluate the system. To reduce inappropriate information, filters were applied to exclude possible confounding factors, such as diseases or symptoms occurring before the use of therapeutic drugs, which were unlikely to have occurred. Overall recall and precision were 75% and 31%, respectively, for identifying known ADRs for the seven drugs. Importantly, the study also suggested that the system has the potential to discover novel ADRs (e.g. 'feeling suicidal'), although the information provided was very limited [10]. Haerian *et al*. used MedLEE with a filter built with expert knowledge and standardized terms. They applied the system to discharge summaries for patients with elevated creatine kinase concentrations (above five times normal), which is known to be associated with rhabdomyolysis resulting from myopathy-inducing medications or various disease states. After removing incidents of rhabdomyolysis resulting from various disease states, the system was able to identify patients with rhabdomyolysis caused by myopathy-inducing medications with a sensitivity of 96.7% and a specificity of 81.4% [19]. Wang *et al*. used MedLEE to determine symptoms related to disease-manifestation and potential ADRs, in discharge summaries of an inpatient EPR. For the purpose of this review, only results related to ADRs are included. Two filters were applied in order to capture ADRs, a filter capturing drugs mentioned only during the course of hospitalization to eliminate medications not given in the hospital and a filter capturing symptoms or diseases only occurring over the course of hospitalization to avoid already present or past symptoms or diseases. Applying filters to capture ADRs, using medications in the course of hospitalization increased recall from 43% to 48% and precision from 16% to 19%, whereas using potential ADRs from the course of hospitalization improved recall from 43% to 54% and precision from 16% to 23%. When both filters were applied, recall and precision were improved from 43% to 75% and precision from 16% to 31% [20].

## *Frequency and types of ADRs identified*

Table 4 summarizes frequencies and types of ADRs identified in the included studies. Honigman *et al*. identified a total of 864 ADRs corresponding to an ADR rate of 5.5 per 100 patients per year. Of all ADRs identified, the most common ADRs belonged to the four system organ classes: skin, central nervous system, gastrointestinal and respiratory system totalling 84% of the ADRs identified in this study. Therapeutic groups causing the most ADRs were antihypertensives, ACE inhibitors, antibiotics and diuretics. Combined, these four groups were associated with 56% of all ADRs identified. Many ADRs, however, were associated with more than one therapeutic group and with more than one medication within therapeutic groups [17].

A total of 204 events, of which 105 were drug related, were identified in the study by Murff *et al*. The most common ADRs were neuropsychiatric (e.g. drug-induced delirium) (30%) and renal failure (24%). The most frequent

## Table 3

Text-mining techniques including performance in included studies to identify ADRs

| References | Text-mining algorithm | Gold standard | Sensitivity*/Specificity** (%) | PPV***/NPV**** (%) | Recall/Precision | % of ADRs |
|---|---|---|---|---|---|---|
| **Honigman *et al.* [17]** | Combination of four search methods:<br>1) Diagnosis codes<br>2) Allergy rules<br>3) Computer event monitoring rules<br>4) Free text searching of electronic visit notes | Manual chart review | Combined search methods:<br>Sensitivity 58% (95% CI 18, 98)<br>Specificity 88% (95% CI 87, 88) | Combined search methods:<br>PPV 7.5% (95% CI,6.5, 8.5)<br>NPV 99.2% (95% CI 95.5, 99.98)<br>By search method (PPVs):<br>1) 2.0% (95% CI 0.6, 4.6)<br>2) 48.6% (95% CI 41.9, 55.3)<br>3) 3.3% (95% CI 2.5, 4.2)<br>4) 7.2% (95% CI 8.8, 7.5) | N/A | Of the four search methods, free text searching of clinic notes accounted for 90.6% of the ADRs detected |
| **Murff *et al.* [18]** | Computerized screening tool using simple keyword queries to search discharge summaries for trigger words representing a broad range of adverse events (AEs) including ADRs.<br>The study population was divided into cohort 1 and cohort 2. Cohort 1 was used to pilot test the screening tool and involved manual chart review of any screened-positive chart as well as manual review of a random 25% sample of screened negative charts. Cohort 2 involved manual chart review of a random 15% sample of all screened – positive charts. | Manual chart review | Cohort 1:<br>Electronic only<br>Sensitivity 69% (95% CI 62, 76)<br>Specificity 48% (95% CI 42, 54)<br>(Based on the screening tool alone)<br>Electronic + manual review<br>Sensitivity 64% (95% CI 61, 67)<br>Specificity 85% (95% CI 80, 90)<br>Cohort 2:<br>Sensitivity (N/A)<br>Specificity (N/A)<br>Measures are based on the identification of AEs in general. Sensitivity and specificity were not reported for ADRs alone. | Cohort 1:<br>Electronic only<br>PPV 52% (95% CI 46, 58)<br>Electronic + manual review<br>PPV 78% (95% CI 71, 85)<br>Cohort 2:<br>Electronic + manual review<br>PPV 84% (95% CI 81, 87)<br>PPVs are based on identification of AEs in general. PPVs were not reported for ADRs alone. | N/A | 52% of the AEs detected represented ADRs (based on results from Cohort 1). |
| **Field *et al.* [21]** | Free text searching using multiple sources:<br>1) Provider reports via an internal ADR reporting system<br>2) Manual review of discharge summaries<br>3) Manual review of notes from emergency department visits<br>4) Computer-generated signals<br>5) Automated free text review of electronic clinic notes<br>6) Manual review of administrative incident reports | Manual chart review | N/A | By source (PPVs):<br>1) 54%<br>2) 5%<br>3) 2%<br>4) 7%<br>5) 12%<br>6) 4% | N/A | By source:<br>1) 11%<br>2) 11%<br>3) 13%<br>4) 31%<br>5) 39%<br>6) 4%<br>Percentages total more than 100% because some ADRs were identified through more than one source |
| **Hazlehurst *et al.* [22]** | Automated classification system (MediClass) using natural language processing and knowledge-based techniques compared to standard methods using diagnosis and utilization codes | Manual chart review | N/A | NLP-based system:<br>PPV 74%<br>Code-based method:<br>PPV 31% | N/A | N/A |
| **Wang *et al.* [10]** | MedLEE natural language processor using discharge summaries | N/A | N/A | N/A | Recall 75%<br>Precision 31% | N/A |
| **Haerian *et al.* [19]** | MedLEE natural language processor using discharge summaries | Manual chart review | Sensitivity 96.7%<br>Specificity 81.4% | N/A | N/A | N/A |
| **Wang *et al.* [20]** | MedLEE natural language processor with regular and contextual filters applied to textual discharge summaries | MedLEE with no filters applied | N/A | N/A | No filters:<br>Recall 43%<br>Precision 16%<br>With filters:<br>Filter 1<br>Recall 48% (95% CI 43, 51)<br>Precision 19% (95% CI 13, 22)<br>Filter 2<br>Recall 54% (95% CI 50, 57)<br>Precision 23% (95% CI 19, 26)<br>Filter 1 + 2<br>Recall 75% (95% CI 68, 79)<br>Precision 31% (95% CI 26, 35) | N/A |

*Sensitivity. The proportion of ADRs, which are correctly identified as such.
**Specificity. The proportion of events not related to medicine use which are correctly identified.
***PPV (positive predictive value). The proportion of events which are correctly identified as ADRs.
****NPV (negative predictive value). The proportion of events which are correctly identified as not being related to medicine use.

## Table 4

Frequency and types of ADRs identified

| Reference | Total ADRs identified (n) | ADR rate (%) | ADRs investigated | Types of ADRs identified (%) | Therapeutic Groups associated with ADRs (%) |
|---|---|---|---|---|---|
| **Honigman *et al.* [17]** | 864 (95% CI, 750, 978) | 5.5% (95% CI, 5.2, 5.9) | N/A | Skin: rash, pedal oedema, angioedema, other (26%)<br>Central nervous system: dizziness, fatigue, other (24%)<br>Gastrointestinal: nausea, diarrhoea, other (18%)<br>Respiratory: cough, shortness of breath (16%)<br>Haematological: overanticoagulation with and without bleeding, other (7%)<br>Cardiovascular: hypotension, tachycardia (4%)<br>Other: N/A (4%) | Antihypertensive: 27.3%<br>ACE inhibitor: 20.7%<br>Antibiotic: 8.3%<br>Diuretic: 7.4%<br>Anticoagulant: 6.6%<br>Antidepressant: 6.6%<br>Cardiovascular: 6.6%<br>Cholesterol-lowering: 5.0%<br>Peptic ulcer: 5.0%<br>NSAID: 5.0%<br>Antiseizure: 3.3% |
| **Murff *et al.* [18]** | 105 (based on results from cohort 1) | N/A | N/A | Cohort 1<br>Neuropsychiatric (e.g. drug-induced delirium) (30%)<br>Renal failure (24%)<br>Dermatological/allergic (13%)<br>Only the most common ADRs were reported<br>Cohort 2<br>N/A<br>The primary outcome of the study was an AE, defined as an injury resulting from medical management rather than the patient's underlying condition. Only results for drug-related AEs are included in this table. | Cohort 1<br>Cardiovascular drugs: 18%<br>Anti-infectives: 13%<br>Anticoagulants: (13%)<br>Cohort 2<br>N/A<br>The primary outcome of the study was an AE, defined as an injury resulting from medical management rather than the patient's underlying condition. Only results for drug-related AEs are included in this table. |
| **Field *et al.* [21]** | 1523 | N/A | N/A | By medication class:<br>Angiotensin-converting enzyme inhibitors: cough<br>Selected antidepressants: anorexia, constipation, hypotension, insomnia, nervousness, dry mouth<br>β-adrenoceptor blockers: bradycardia<br>Hypoglycaemics: hypoglycaemia, tremor<br>NSAIDs: bleeding, gastrointestinal complaints, nausea, renal failure/insufficiency<br>Diuretics: hyponatremia, hypotension<br>Warfarin: bleeding<br>Digoxin: nausea<br>Opioids: constipation<br>Calcium channel blockers: peripheral oedema<br>Proton pump inhibitors: diarrhoea<br>Antibiotics: diarrhoea | N/A |
| **Hazlehurst *et al.* [22]** | N/A | N/A | Vaccine related ADRs (in particular gastrointestinal) | N/A | N/A |
| **Wang *et al.* [10]** | 132 | N/A | Known ADRs associated with ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone, ACE inhibitors | By medication class:<br>Ibuprofen: headache, achalasia, nausea, constipation<br>Bupropion: dizziness, abnormal sensation, difficulty, drugged state, fatigue, constipation, sleepiness, seizure, tinnitus, pruritus, *feeling suicidal*<br>Paroxetine: chest pain, drowsiness, orthostasis, dyspnoea, agitation, dizziness, *feeling suicidal*<br>Rosiglitazone: headache, *chest pain, left atrial hypertrophy, shortness of breath*<br>ACE-inhibitors: cough, lethargy, dizziness, diarrhoea, headache<br>Only five out of seven medication classes were listed in the article due to space limitations.<br>ADRs in italics were emphasized by the authors as being novel. | N/A |
| **Haerian *et al.* [19]** | 165 | N/A | Rhabdomyoly-sis resulting from myopathy inducing medications | Rhabdomyolysis | Valproic acid<br>Tenofovir<br>Statins<br>Haloperidol |
| **Wang *et al.* [20]** | N/A | N/A | Drug–adverse drug event relation (not further specified) | N/A | N/A |

therapeutic groups associated with ADRs were cardiovascular drugs (18%), anti-infectives (13%) and anticoagulants (13%). These figures are based on results from cohort 1 only (see Table 3) [18].

Field *et al*. included a total of 381 249 patients in their study examining strategies for detecting ADRs among older persons in an ambulatory setting. A total of 1523 ADRs were identified during the 1 year tracking period. The most commonly detected ADR–drug associations were cough and angiotensin converting enzyme, hypoglycaemia and hypoglycaemics, constipation and opioids, peripheral oedema and calcium channel blockers and diarrhoea from the use of antibiotics. As multiple sources (see Table 3) were included in the study to identify possible ADRs, it is not possible to describe frequency by type of ADRs [21].

Wang *et al*. detected 132 ADRs associated with the seven medications, ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone and ACE inhibitors. Information related to the frequency of ADRs by medication class, however, was not available and specific types of ADRs were only listed for five of the seven drugs due to space limitations (Table 4). The study also suggested that novel ADRs could be detected, including feeling suicidal associated with bupropion and paroxetine, chest pain, left atrial hypertrophy and shortness of breath associated with rosiglitazone [10].

In their study on rhabdomyolysis, Haerian *et al*. included 687 inpatients with elevated creatine kinase. In a total of 165 patients, the elevated creatine kinase was attributed to myopathy-inducing medications. Several medications known to cause an elevated creatine kinase or rhabdomyolysis were revealed including valproic acid, tenofovir, statins and haloperidol [19].

Two studies [20, 22] did not provide sufficient information to allow description of frequency and specific types of ADRs identified.

## Discussion

In this review we show that very few published studies have used text-mining of EPRs to investigate ADRs and that initial studies used relatively simple text-searching techniques whereas more recent studies have focused on more advanced techniques such as NLP. Although performance appears to have increased with the use of NLP-based techniques, many ADRs were missed compared with manual chart review. The studies did not provide sufficient information regarding frequency and types of ADRs identified to allow comparisons across studies.

Despite the very few studies included in this review, our findings indicate that text mining of EPRs is a feasible and potentially promising means of identifying ADRs. Electronically screening narrative EPR documents for ADRs is advantageous in that it requires less time than manual chart review and thus is cheaper. The time gains acquired from electronic screening in the included studies, however, are limited by poor specificity and low PPVs. At present, text mining cannot entirely replace manual chart review, but it can be a valuable supplement, in particular when screening large amounts of text.

Text mining of narrative EPR documents is efficient compared with several other computerized methods used for detecting ADRs. Two of the studies included in the analysis compared various search methods and both found that free text searching of electronic clinic notes accounted for the highest number of ADRs detected [17, 21]. In the study by Hazlehurst *et al*. the NLP-based system identified approximately four times as many ADRs compared with standard methods that use diagnosis and utilization codes [22].

The approaches adopted by the included studies to detect information about ADRs in EPRs have varied over time from use of simple free text searching of outpatient visit notes and inpatient discharge summaries to more advanced techniques involving NLP of inpatient discharge summaries. Simple free text searching is practical because it requires only that dictated text is available electronically. Nonetheless, this computerized method requires major adjustments including eliminating sentences with negative terms or ambiguous terms [17, 18]. Even after such modifications, simple text searching methods in the included studies yielded relatively low PPVs [17, 21]. Performance appears to increase with the use of NLP-based techniques [10, 19, 20, 22]. However, the complexity of the study designs and use of filters make it difficult to conclude whether NLP did in fact contribute to the increased performance. Another unsolved question is whether the use of alternative terminologies for coding of ADRs (e.g. SNOMED-CT) can improve the detection rate of ADRs [23].

The articles generally provided too little information to allow a detailed description of frequency and types of ADRs identified. Furthermore, the studies used different approaches to identify ADRs. Four studies investigated ADRs in general [17, 18, 20, 21], while the remaining three studies investigated specific types of ADRs (vaccine-related ADRs) [22], known ADRs associated with selected medications [10] and rhabdomyolysis [19]. Importantly, only one study emphasized the detection of new ADRs [10], while the remaining studies identified known ADRs. If text-mining narrative EPR documents for identification of ADRs is to revolutionize pharmacovigilance, this underscores the need for new empirical studies with focus on ADR detection (in particular new, rare and long term ADRs) rather than text-mining methodology.

### *Strengths and limitations of the study*

Text mining narrative EPR data to investigate ADRs is a relatively new and promising field of research and a critical review is therefore highly relevant to summarize what approaches have been explored to date and what types of

ADRs can be identified. We searched a total of four literature databases covering international journals within biomedicine and pharmacology without restrictions and double-checked references. Nevertheless, we were able to identify only seven articles for inclusion in the review, published from 2001 and onwards.

The review has several limitations. We used the EU definition of an ADR. Many of the studies, however, did not clearly define ADRs or focused on the identification of adverse events in general, including drug-related adverse events. Furthermore, we question the validity of the ADRs identified by the studies, as manual chart review was only performed on a small proportion of the charts. Despite the titles of the articles, the focus of the studies was on text-mining methodology rather than on identification of ADRs and information in relation to frequency and types of ADRs was, therefore, generally limited. Furthermore, conflicting study designs and populations resulted in various different ADRs. Hence, making direct comparisons across studies with regard to frequency and types of ADRs was not possible. Despite the limitations, the review confirms that the application of text mining has potential for identifying ADRs in EPRs, although substantial work remains before text mining of EPRs can be applied for drug surveillance.

## Conclusion

This review underscores the feasibility and potential of text mining to investigate narrative documents in EPRs for the presence of ADRs. Due to poor specificity and low PPVs of current text-mining tools, however, text mining cannot entirely replace manual chart review but can be a supplement to screening of large amounts of data. Additionally, more empirical studies should evaluate whether text mining of EPRs can be used systematically to collect new information about ADRs.

## Competing Interests

None of the authors had any competing interests to declare regarding the content of this article.

*This study was part of a Ph. D. programme financed by the Faculty of Pharmaceutical Sciences, University of Copenhagen, Denmark.*

## Appendix 1 Search strategy for database searches from origin to July 2011

### *PubMed*

**#5** Search **#1** AND **#2** AND **#3** AND **#4**

**#4** 'mining' [MeSH] OR 'data mining' [MeSH] AND 'drug toxicity' [MeSH] AND 'medical records systems, computerized' [MeSH]

**#3** 'clinical record* [keyword] OR 'computerized medical record*' [keyword] OR 'computerized patient record*' [keyword] OR 'electronic health record*' [keyword] OR 'electronic medical record*' [keyword] OR 'electronic medical record*' [keyword] OR 'free-text medical record*' [keyword] OR 'medical record* [keyword] OR narrative patient record* [keyword] OR *patient record system* [keyword] OR 'unstructured textual information*' [keyword]

**#2** 'adverse drug event*' [keyword] OR 'adverse drug reaction*' [keyword] OR 'adverse event*' [keyword] OR 'adverse reaction*' [keyword] OR 'drug toxicity' [keyword] OR 'pharmacovilance' [keyword]

**#1** 'information extraction' [keyword] OR 'knowledge acquisition' [keyword] OR 'knowledge extraction' [keyword] OR 'mining' [keyword] OR 'natural language processing' OR 'text mining' [keyword]

### *Embase*

#4 Search #1 and #2 and #3

#3 (clinical record* or computerized medical record* or computerized patient record* or electronic health record* or electronic medical record* or electronic patient record* or free-text medical record* or medical record*or medical records system* or narrative patient record* or patient record system* or unstructured textual information*).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer]

#2 (adverse drug event* or adverse drug reaction* or adverse event* or adverse reaction* or drug toxicit* or pharmacovigilance).mp. [mp = title, abstract, subject headings, heading word, drug trade name, original title, device manufacturer, drug manufacturer]

#1 (data mining or information extraction or knowledge acquisition or knowledge extraction or mining or natural language processing or text mining)

### *Web of science*

#4 Search #1 and #2 and #3

Databases = SCI-EXPANDED, SSCI, A&HCI Timespan = All Years

#3 Topic = (clinical record*) OR Topic = (computerized medical record*) OR Topic = (computerized patient record*) OR Topic = (electronic health record*) OR Topic = (electronic medical record*) OR Topic = (medical record*) OR Topic = (medical records system) OR Topic = (narrative patient record*) OR Topic = (patient record system*) OR Topic = (unstructured textual information)

Databases = SCI-EXPANDED, SSCI, A & HCI Timespan = All Years

#2 Topic = (adverse drug event*) OR Topic = (adverse drug reaction*) OR Topic = (adverse event*) OR Topic = (adverse reaction*) OR Topic = (drug toxit*) OR Topic = (pharmacovigilance)

Databases = SCI-EXPANDED, SSCI, A & HCI Timespan = All Years

#1 Topic = (data mining) OR Topic = (information extraction) OR Topic = (knowledge acquisition) OR Topic = (knowledge extraction) OR Topic = (mining) OR Topic = (natural language processing) OR Topic = (text mining)

*Databases = SCI-EXPANDED, SSCI, A&HCI Timespan = All Years*

### *IPA*

#4 Search #1 and #2 and #3

#3 (clinical record$ or computerized medical record$ or computerized patient record$ or electronic health record$ or electronic medical record$ or electronic patient record$ or free-text medical record$ or medical record$ or medical records system$ or narrative patient record$ or unstructured textual information).mp. [mp = title, subject heading word, registry word, abstract, trade name/generic name]

#2 (adverse drug event$ or adverse drug reaction$ or adverse event$ or adverse reaction$ or drug toxicit$ or pharmacovigilance).mp. [mp = title, subject heading word, registry word, abstract, trade name/generic name]

#1 (information extraction$ or knowledge acquisition$ or knowledge extraction$ or mining$ or natural language processing$ or text mining$ or data mining$).mp. [mp = title, subject heading word, registry word, abstract, trade name/generic name]

## REFERENCES


**1** Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, Farrar K, Park BK, Breckenridge AM. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18.820 patients. BMJ 2004; 329: 15–9.

**2** Patel P, Zed PJ. Drug-related visits to the emergency department: how big is the problem? Pharmacotherapy 2002; 22: 915–23.

**3** Wester K, Jonsson AK, Spigset O, Druid H, Hagg S. Incidence of fatal adverse drug reactions: a population based study. Br J Clin Pharmacol 2008; 65: 573–9.

**4** Hansen EH, Launsø L. Is the controlled clinical trial sufficient as a drug technology assessment? J Soc Adm Pharm 1989; 6: 117–26.

**5** Stricker BH, Psaty BM. Detection, verification, and quantification of adverse drug reaction. BMJ 2004; 3: 44–7.

**6** Aagaard L, Soendergaard B, Andersen E, Kampmann JP, Hansen EH. Creating knowledge about adverse drug reactions: a critical analysis of the Danish reporting system from 1968 to 2005. Soc Sci Med 2007; 65: 1296–309.

**7** Aagaard L, Hansen EH. Information about ADRs explored by pharmacovigilance approaches: a qualitative review of studies on antibiotics, SSRIs and NSAIDs. BMC Clin Pharmacol 2009; 9: 4.

**8** Cerrito P. Application of data mining for examining polypharmacy and adverse effects in cardiology patients. Cardiovasc Toxicol 2001; 1: 177–9.

**9** Glascow JM, Kaboli PJ. Detecting adverse drug events through data mining. Am J Health Syst Pharm 2010; 67: 317–20.

**10** Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009; 16: 328–37.

**11** Wikipedia. Text mining. 2011. Available at http://en.wikipedia.org/org/wiki/Textmining (last accessed 9 August 2011).

**12** Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000; 33: 1–10.

**13** Sager N, Lyman M, Tick LJ, Nhan NT, Bucknall CE. Natural language processing of asthma discharge summaries for the monitoring of patient care. Proc Annu Symp Comput Appl Med Care 1993; 265–8.

**14** Hazlehurts B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, Winickoff JP, Glasgow R, Palen TE, Rigoffi NA. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med 2005; 29: 434–9.

**15** Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 2011; 7: art.no. e102141.

**16** EudraLex. Volume 9. Pharmacovigilance: medicinal products for human use and veterinary products. 2011. Available at http://ec.europa.eu/health /documents/eudralex/vol-9/index en.htm (last accessed 9 August 2011).

**17** Honigman B, Lee J, Rothschild J, Light P, Pulling RM, Yu T, Bates DW. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc 2001; 8: 254–66.

**18** Murff HJ, Forster AJ, Peterson JF, Fiskio JM, Heiman HL, Bates DW. Electronically screening discharge summaries for adverse medical events. J Am Med Inform Assoc 2003; 10: 339–50.

**19** Haerian K, Varn D, Chase H, Vaidya S, Friedman C. Electronic health record pharmacovigilance signal extraction: a semi-automated method for reduction of confounding applied to detection of rhabdomyolysis. 10th Annual Meeting of the International Society of Pharmacovigilance. ISoP 2010; 33: 933.

**20** Wang X, Chase H, Markatou M, Hripcsak G, Friedman C. Selecting information in electronic health records for knowledge acquisition. J Biomed Inform 2010; 43: 595–601.

**21** Field TS, Gurwitz JH, Harrold LR, Rothschild JM, Debellis K, Seger AC, Fish LS, Garber L, Kelleher M, Bates DW. Strategies for detecting adverse drug events among older persons in the ambulatory setting. J Am Med Inform Assoc 2004; 11: 492–8.

**22** Hazlehurst B, Naleway A, Mulooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. Vaccine 2009; 23: 2077–83.

**23** Simpson CR, Anandan C, Fischbacher C, Lefevre K, Sheikh A. Will systematized Nomenclature of Medicine – Clinical Terms improve our understanding of the disease burden posed by allergic disorders? Clin Exp Allergy 2007; 37: 1586–93.