Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics

Semi-automatische Deidentifizierung von deutschsprachigen medizinischen Berichten mit vertraulichem Inhalt für Big Data Analysen

Authors
Hannes Seuss1, Peter Danker1, Matthias Ihle2, Andrea Grandjean2, Rebecca Hammon2, Nicola Kaestle3, Peter A. Fasching5, Christian Maier4, Jan Christoph4, Martin Sedlmayr6, Michael Uder1, Alexander Cavallaro1, Matthias Hammon1

Affiliations
1 Department of Radiology, University Hospital Erlangen, Friedrich Alexander Universität (FAU) Erlangen-Nürnberg, Erlangen, Germany
2 Text Analytics, Averbis GmbH, Freiburg, Germany
3 Department of Neurology, Klinikum Nuremberg, Nuremberg, Germany
4 Department of Neuroradiology, University Hospital Erlangen, Erlangen, Germany
5 Department of Gynecology and Obstetrics, Comprehensive Cancer Center Erlangen-EMN, Erlangen University Hospital, Friedrich Alexander University of Erlangen-Nuremberg, Erlangen, Germany
6 Lehrstuhl für Medizinische Informatik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Key words
big data, data mining, de-identification, anonymization, data scrubbing, medical reports

Bibliography
DOI http://dx.doi.org/10.1055/s-0043-102939
Published online: 23.3.2017 | Fortschr Röntgenstr 2017; 189: 661–671 © Georg Thieme Verlag KG Stuttgart · New York
ISSN 1438-9029

Correspondence
PD Dr. Matthias Hammon
Department of Radiology, Universitätsklinikum Erlangen, Friedrich Alexander Universität (FAU) Erlangen-Nürnberg Maximilianplatz 1 91054 Erlangen Germany
Tel.: ++49/91 31/8 53 60 65
Fax: ++49/91 31/8 53 60 68
matthias.hammon@uk-erlangen.de

ZUSAMMENFASSUNG
Ziel

Material und Methoden
Die Software für Text-Mining und Deidentifizierung wurde in medizinischen Berichten zur Erfassung sensibler Inhalte auf ihre Sensitivität und Spezifität getestet. 4671 pathologische (4105 + 566 in zwei unterschiedlichen Formaten), 2804 medizinische, 1008 operative und 6223 radiologische Berichte von 1167 Patientinnen und Patienten, die an Brustkrebs leiden, wurden deidentifiziert. Der Inhalt wurde in vier Kategorien aufgeschlüsselt: direkte Kennung (Name, Adresse), indirekte Kennung (Geburtsdatum, Operationsdatum, medizinische ID, etc.), medizinische Begriffe und Füllwörter. Die Software wurde nativ getestet (ohne Training), um einen Ausgangswert zu erhalten. Anschließend wurde das Modell an manuell korrigierten Berichten erneut trainiert. Nach der Bearbeitung von 25, 50, 100, 250, 500 und 1000 Berichten eines jeden Typs, wurde ein erneutes Training durchgeführt.

Ergebnisse
Nativ wurden 61,3 % der direkten und 80,8 % der indirekten Kennungen nachgewiesen. Nach dem Training erhöhte sich die Leistung (P) auf 91,4 % (P25), 96,7 % (P50), 99,5 % (P100), 99,6 % (P250), 99,7 % (P500) und 100 % (P1000) für direkte Kennungen und 93,2 % (P25), 97,9 % (P50), 97,2 % (P100), 98,9 % (P250), 99,0 % (P500) und 99,3 % (P1000) für indirekte Kennungen. Im Durchschnitt wurden 5,3 % der medizinischen Begriffe als kritische Daten gekennzeichnet, nach dem Training waren es 4,0 % (P25), 3,6 % (P50), 4,0 % (P100), 3,7 % (P250), 4,3 % (P500), 3,1 % (P1000). Etwa 0,1 % der Füllwörter wurden gekennzeichnet.

Schlussfolgerung
Das Training der entwickelten Deidentifikations-Software verbessert ihre Performance kontinuierlich. Das Training mit etwa 100 korrigierten Texten ermöglicht eine zuverlässige Detektion und Markierung der sensiblen Daten in unterschiedlichen medizinischen Texten.

Kernaussagen:
- Wenn Patientendaten zwischen unterschiedlichen Institutionen ausgetauscht werden, müssen diese zuvor deidentifiziert werden

**ABSTRACT**

**Purpose** Projects involving collaborations between different institutions require data security via selective de-identification of words or phrases. A semi-automated de-identification tool was developed and evaluated on different types of medical reports natively and after adapting the algorithm to the text structure.

**Materials and Methods** A semi-automated de-identification tool was developed and evaluated for its sensitivity and specificity in detecting sensitive content in written reports. Data from 4671 pathology reports (4105 + 566 in two different formats), 2804 medical reports, 1008 operation reports, and 6223 radiology reports of 1167 patients suffering from breast cancer were de-identified. The content was itemized into four categories: direct identifiers (name, address), indirect identifiers (date of birth/operation, medical ID, etc.), medical terms, and filler words. The software was tested natively (without training) in order to establish a baseline. The reports were manually edited and the model re-trained for the next test set. After manually editing 25, 50, 100, 250, 500 and if applicable 1000 reports of each type re-training was applied.

**Results** In the native test, 61.3% of direct and 80.8% of the indirect identifiers were detected. The performance (P) increased to 91.4% (P25), 96.7% (P50), 99.5% (P100), 99.6% (P250), 99.7% (P500) and 100% (P1000) for direct identifiers and to 93.2% (P25), 97.9% (P50), 97.2% (P100), 98.9% (P250), 99.0% (P500) and 99.3% (P1000) for indirect identifiers. Without training, 5.3% of medical terms were falsely flagged as critical data. The performance increased, after training, to 4.0% (P25), 3.6% (P50), 4.0% (P100), 3.7% (P250), 4.3% (P500), and 3.1% (P1000). Roughly 0.1% of filler words were falsely flagged.

**Conclusion** Training of the developed de-identification tool continuously improved its performance. Training with roughly 100 edited reports enables reliable detection and labeling of sensitive data in different types of medical reports.

**Key Points:**
- Collaborations between different institutions require de-identification of patients’ data
- Software-based de-identification of content-sensitive reports grows in importance as a result of ‘Big data’
- A de-identification software was developed and tested natively and after training
- The proposed de-identification software worked quite reliably, following training with roughly 100 edited reports
- A final check of the texts by an authorized person remains necessary

**Citation Format**

**Introduction**

“What I may see or hear in the course of the treatment or even outside of the treatment in regard to the life of men, which on no account one must spread abroad, I will keep to myself, holding such things shameful to be spoken about” [1]. This part of the Hippocratic Oath laid the groundwork for today’s patient information confidentiality. Many national and international laws like the “Health Insurance Portability and Accountability Act” (HIPAA), the “Common Rule”, and “Directive 2002/58/EC” of the European Parliament and of the Council were formalized in accordance with this oath [2–4].

While keeping patient data safe, scientific collaboration between different (external) institutions is a vital part of today’s medical research. For example, in a study evaluating a new drug, several disciplines have to work together and exchange information about the treatment of the patient. While all participants are involved in the treatment of the patient, there will be no issue with doctor-patient confidentiality and the laws governing data protection. However, as soon as external partners like pharmaceutical companies become involved, medical texts have to be de-identified to secure the patient’s privacy [5].

It is necessary to distinguish between de-identification or pseudonymization and anonymization [6]. De-identification removes or replaces all personal identifiers, but authorized individuals are still able to relink the data to the patient, usually via a hash table. In an anonymized report, the link to the patient is irreversibly lost and it is virtually impossible to connect the record to the patient.

An identifier is every piece of information that can be used to identify a person. The most obvious is the name, followed by address, and social security number. However, information that seems harmless at first can be combined with publically available data to directly identify a person. How easy an individual can be identified was shown by Sweeney who purchased the 1997 voting list of Cambridge Massachusetts and was able to uniquely identify 69% of voters only by their birth date and five-digit zip code [7]. HIPAA clearly defines the categories and content of protected health information (Table 1). Important data for de-identification are patient identifiers like name, street address, city, county, zip code, dates, telephone numbers, e-mail addresses, social security numbers, account and medical record numbers, biometric identifiers (laboratory data, genetic code), and any other potential identifier of the patient [8]. The patient’s privacy must be protec-
de-identification is text mining. The unstructured content of plain text is analyzed for key words and grammar, the identifiers are annotated, and in a last step the tags can be blackened or substituted by a hash identifier (ID).

The aim of this study is to evaluate new software that semi-automatically de-identifies information found in different types of medical reports by comparing the accuracy of the software natively and after training.

**Materials and Methods**

**Patient characteristics**

The reports of 1167 patients with histologically confirmed breast cancer were exported retrospectively. The data was further classified by gender: 1153 patients were female and 14 were male. The mean age of patients was 51.4 years (19 to 94 years). The mean duration of case history was 5.3 years. 26 patients tested positive for a BRCA-1 mutation and 30 patients for a BRCA-2 mutation. No known mutation was found in 203 patients and 908 patients were not tested.

This retrospective study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of the University Hospital Erlangen. The need for written informed consent was waived by the Ethics Committee.

**Data characteristics**

A total of 14,706 written reports were exported for the collaboration. The reports consist of 4,671 pathology reports, 2,804 medical reports, 1,008 operative reports and 6,223 radiology reports. The pathology reports came from the Institute of Pathology of the University Hospital Erlangen (UHE) and existed in two different formats, 4,105 in plain text and 566 in the extensible markup language (XML) format. The medical reports (XML) and the operative reports (plain text) were provided by the Department of Obstetrics and Gynecology of the UHE. The radiology reports (plain text) were provided by the Department of Radiology of the UHE. All reports were written in German.
De-identification software

The de-identification process used dedicated software that was developed for semi-automatic de-identification of unstructured clinical records (deID, Averbis GmbH, Freiburg, Germany). The software is a database-driven client-server web application. Original documents are initially imported into the applications database, subsequently displayed and annotated in the web-based graphical user interface (GUI), and finally exported as a de-identified version into a target format.

Several data formats are supported for the import, including plain text and different HL-7 message formats. Support for XML-based formats, including CDA (Clinical Document Architecture), is currently under development.

During annotation, text passages are allocated into categories either manually or automatically. The manual process is engaged in the browser by selecting the passage and choosing the category with the mouse. The automatic process is accomplished by means of the integrated Natural Language Processing (NLP) pipeline presented later on. The predefined categories are "name", "date", "location", "contact", "division", "ID", "age", "biometrics" and "other." Usually, both approaches are combined.

Documents are pre-annotated by the NLP pipeline and manually revised by adding missing and deleting incorrect annotations. The system’s machine learning (ML) component learns and trains using these revised documents. The learning process is used to improve the quality of the automatic annotations, with the goal of rendering manual revision unnecessary after a sufficient number of training iterations.

During the export, de-identified versions of all approved documents are generated, whereby annotated passages are replaced according to a configurable replacement strategy, e.g., by replacing them with string placeholders like "Name", "Date", or "XXX".

De-identification process

The software supports a de-identification process, developed within the scope of the BMWi-funded project cloud4health in coordination with German data protection supervisors. Each supervisor supports different roles such as annotators, administrators and those responsible for data approval. The core of this de-identification process is the repeated training and evaluation of a model, which is generated and used by the machine learning component of the NLP pipeline to recognize annotation patterns from annotated records, and subsequently apply them during automatic annotation. This process is shown in Fig. 1.

Auto-generated annotations

The auto-generated annotations use an NLP pipeline with a multitude of pipeline components. In addition to preprocessing and cleanup, these components can be divided into three high level steps:

- Metadata matching is performed with annotation of the patient’s name, contact information, date of birth, etc. when found in the document text. The availability of such metadata depends on the import document format, e.g., HL7 contains patient metadata in the header while plain text documents lack such additional information.
- The second step is a pre-assigned sequence of different rule-based methods. These methods are simple lexicographic and pattern-based approaches but also a combination of both. In the first case, the given text is tested on expressions from word lists, for example: the existence of location names. The second type of analysis uses predefined patterns, e.g., for phone numbers or dates, and tries to annotate text sequences matching these patterns. The third and even more complex approach combines the previous methods and follows predefined rules in order to find additional text snippets for annotation.
- The machine learning component is based on Conditional Random Fields with standard Named Entity Recognition features (NER features); some additional features are tailored to the characteristics of the clinical records [9]. It models a document as a sequence of words where each word is thereby assigned to one of the annotation categories described in the last chapter.

Re-training the model

In order to adjust the software to the specific structure and content of the text, an additional statistical learning algorithm was included in the NLP pipeline. It was trained using manually corrected and approved reports. Reviews were done by a fourth year radiological resident and two board-certified radiologists.

Evaluation

During the evaluation of the de-identification method, the annotations were itemized into four different categories. The names and addresses of patients or employees are direct identifiers and are considered highly sensitive data in terms of data protection law. Dates (e.g., date of birth, date of operation/examination, dates in patient history), identification numbers or names of studies are indirect identifiers and were still considered critical data. However, further information was needed (access to a database is required) to identify the patient or employee. The annotations could be true positive (TP). If the critical term was correctly marked by the program, it would be labeled TP. If the term was not annotated, it would be labeled false negative (FN). The data which doesn’t have to be labeled is itemized into appropriate categories such as: important medical content, i.e., “heart” or “liver”; negations, i.e., “no” or “not”; unimportant filler words, i.e., most of the verbs; or phrases like “please” or “for” that did not alter information. Data that should be unlabeled could be false positive (FP), therefore labeled as critical or true negative (TN), therefore, correctly kept in the report. TP, FP and FN were recorded manually. A word count function was integrated into the program. The number of filler words was calculated by subtracting the findings of other categories from the total word count.

The performance (P) of the program was evaluated without training (P0) and after re-training with 25 (P25), 50 (P50), 100 (P100), 250 (P250), 500 (P500) and if applicable 1000 reports (P1000) for the different types of reports. P0 was evaluated on
the first 25 reports, P25 on reports 26–50, and P50 on reports 51–100. Therefore, all reports of the training data set were analyzed. For the evaluation of P100 to P1000, the test data set consisted of only the first 50 reports of each training data set.

Results

Of the 14706 manually corrected reports, a total of 1400 documents with 500921 words were analyzed for this study; of these, 4563 words were direct and 18976 indirect identifiers; 93199 words were of medical importance; and 4884183 were filler words. The following results are the summary for operative, pathology, radiology and medical reports. Detailed information for the separate groups is shown in ▶ Table 2, 3 and in ▶ Fig. 2.

4299 (94.2%) of the direct identifiers were flagged correctly. Without training (P0) 61.3% of the direct identifiers were labeled correctly. The sensitivity improved to 91.4% (P25), 96.7% (P50), 99.5% (P100), 99.6% (P250), 99.7% (P500), and 100% (P1000) after training with different numbers of texts.

18286 (96.4%) of the indirect identifiers were flagged correctly. Without training (P0), 80.8% of the indirect identifiers were flagged correctly. The sensitivity improved to 93.2% (P25), 97.9% (P50), 97.2% (P100), 98.9% (P250), 99.0% (P500), and 99.3% (P1000) with different numbers of training texts.

Without training, the identification of important medical terms failed in 5.3% (P0). After training with different numbers of texts, the results were 4.0% (P25), 3.6% (P50), 4.0% (P100), 3.7% (P250), 4.3% (P500), and 3.1% (P1000). 0.2% of filler words were falsely flagged.

Without itemization into groups, the total sensitivity was 76.5% (P0), 92.8% (P25), 97.6% (P50), 97.6% (P100), 99.1% (P250), 99.1% (P500), 99.4% (P1000), the total specificity 99.8% (P0), 99.4% (P25), 99.4% (P50), 99.9% (P100), 99.8% (P250), 99.8% (P500), 99.9% (P1000), the positive predictive value 67.3% (P0), 41.2% (P25), 49.4% (P50), 78.2% (P100), 73.5%
### Table 2

Results of the accuracy of the de-identification software. For direct and indirect identifiers, the true positives, false negatives, total tokens and the sensitivity are shown for the different kinds of reports. For medical terms and filler words the true negatives, false positives, total tokens and the specificity are shown.

<table>
<thead>
<tr>
<th></th>
<th>direct identifier</th>
<th>indirect identifier</th>
<th>medical terms</th>
<th>filler words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FN</td>
<td>total</td>
<td>sensitivity</td>
</tr>
<tr>
<td>1008 operation reports (n = 250)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>202</td>
<td>61</td>
<td>263</td>
<td>76.8</td>
</tr>
<tr>
<td>50</td>
<td>480</td>
<td>6</td>
<td>486</td>
<td>98.8</td>
</tr>
<tr>
<td>100</td>
<td>354</td>
<td>1</td>
<td>355</td>
<td>99.7</td>
</tr>
<tr>
<td>250</td>
<td>359</td>
<td>1</td>
<td>360</td>
<td>99.7</td>
</tr>
<tr>
<td>500</td>
<td>323</td>
<td>0</td>
<td>323</td>
<td>100.0</td>
</tr>
<tr>
<td>566 pathology reports (XML) (n = 250)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>59</td>
<td>21</td>
<td>80</td>
<td>73.8</td>
</tr>
<tr>
<td>50</td>
<td>166</td>
<td>13</td>
<td>179</td>
<td>92.7</td>
</tr>
<tr>
<td>100</td>
<td>158</td>
<td>2</td>
<td>160</td>
<td>98.8</td>
</tr>
<tr>
<td>250</td>
<td>172</td>
<td>1</td>
<td>173</td>
<td>99.5</td>
</tr>
<tr>
<td>500</td>
<td>138</td>
<td>1</td>
<td>139</td>
<td>99.3</td>
</tr>
<tr>
<td>4105 pathology reports (n = 300)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>58</td>
<td>0</td>
<td>58</td>
<td>100.0</td>
</tr>
<tr>
<td>50</td>
<td>101</td>
<td>0</td>
<td>101</td>
<td>100.0</td>
</tr>
<tr>
<td>100</td>
<td>87</td>
<td>0</td>
<td>87</td>
<td>100.0</td>
</tr>
<tr>
<td>250</td>
<td>65</td>
<td>0</td>
<td>65</td>
<td>100.0</td>
</tr>
<tr>
<td>500</td>
<td>67</td>
<td>0</td>
<td>67</td>
<td>100.0</td>
</tr>
<tr>
<td>1000</td>
<td>57</td>
<td>0</td>
<td>57</td>
<td>100.0</td>
</tr>
<tr>
<td>6223 radiology reports (n = 300)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>15</td>
<td>1</td>
<td>16</td>
<td>93.8</td>
</tr>
<tr>
<td>50</td>
<td>27</td>
<td>1</td>
<td>28</td>
<td>96.4</td>
</tr>
<tr>
<td>100</td>
<td>34</td>
<td>0</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>250</td>
<td>32</td>
<td>0</td>
<td>32</td>
<td>100.0</td>
</tr>
<tr>
<td>500</td>
<td>34</td>
<td>0</td>
<td>34</td>
<td>100.0</td>
</tr>
<tr>
<td>1000</td>
<td>37</td>
<td>0</td>
<td>37</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**Table 2 (Continuation)**

<table>
<thead>
<tr>
<th></th>
<th>direct identifier</th>
<th>indirect identifier</th>
<th>medical terms</th>
<th>filler words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FN</td>
<td>total</td>
<td>sensitivity</td>
</tr>
<tr>
<td>2804 medical reports (n = 300)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>45</td>
<td>61</td>
<td>106</td>
<td>42.5 %</td>
</tr>
<tr>
<td>25</td>
<td>73</td>
<td>19</td>
<td>92</td>
<td>79.3 %</td>
</tr>
<tr>
<td>50</td>
<td>153</td>
<td>12</td>
<td>165</td>
<td>92.7 %</td>
</tr>
<tr>
<td>100</td>
<td>186</td>
<td>1</td>
<td>187</td>
<td>99.5 %</td>
</tr>
<tr>
<td>250</td>
<td>175</td>
<td>1</td>
<td>176</td>
<td>99.4 %</td>
</tr>
<tr>
<td>500</td>
<td>184</td>
<td>1</td>
<td>185</td>
<td>99.5 %</td>
</tr>
<tr>
<td>1000</td>
<td>187</td>
<td>0</td>
<td>187</td>
<td>100.0 %</td>
</tr>
<tr>
<td>all 14 706 reports (n = 1400)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>288</td>
<td>182</td>
<td>470</td>
<td>61.3 %</td>
</tr>
<tr>
<td>25</td>
<td>435</td>
<td>41</td>
<td>476</td>
<td>91.4 %</td>
</tr>
<tr>
<td>50</td>
<td>927</td>
<td>32</td>
<td>959</td>
<td>96.7 %</td>
</tr>
<tr>
<td>100</td>
<td>819</td>
<td>4</td>
<td>823</td>
<td>99.5 %</td>
</tr>
<tr>
<td>250</td>
<td>803</td>
<td>3</td>
<td>806</td>
<td>99.6 %</td>
</tr>
<tr>
<td>500</td>
<td>746</td>
<td>2</td>
<td>748</td>
<td>99.7 %</td>
</tr>
<tr>
<td>1000</td>
<td>281</td>
<td>0</td>
<td>281</td>
<td>100.0 %</td>
</tr>
</tbody>
</table>
The five medical terms that were falsely flagged most often were: 99 times “Herz” (“heart”), 81 times “Winkel” (“angle”), 59 times “ED MaCa” (“initial diagnosis of breast cancer”), 25 times “Leber” (“liver”) and 24 times “Milz” (“spleen”).

**Discussion**

Working with medical reports is a balancing act between sufficiently protecting the patient’s privacy and efficiently working with the data. If external institutions are involved, reports must be de-identified reliably. Therefore, the aim of this study was to evaluate semi-automated de-identification software on different types of medical reports natively and following several steps of training with manually edited reports. Training of the proposed tool continuously improved its performance. Training with roughly 100 edited reports enables quite reliable detection and labeling of sensitive data in different types of medical reports.

It was shown that it is important to not only erase or replace the name of a patient in a medical report, but to completely strip the report of any information that could link patient and report [7]. In the United States of America HIPAA clearly states what kind of information is critical to identify a person. However, it is not enough to remove the name and address from the report. If the patient has a rare disease, even the medical information in the report can be used to identify the patient.

Many studies analyzing the task of de-identification have been conducted. Kushida et al. reviewed 34 articles dealing with strategies for de-identifying or anonymizing written medical reports [6]. They concluded that current de-identification strategies have their limitations, and statistical learning-based systems have distinct advantages over other approaches for the de-identification of free text.

Thomas et al. used an augmented “search and replace” method for de-identifying pathology reports [10]. They took advantage of the fact that most proper names in their examined report occurred in pairs. They only missed 3 of 231 names and therefore were able to de-identify 98.7 % of proper names in the prose section. This system was limited to patient’s names; no birthdates, addresses, or IDs were analyzed by the program.

Gupta et al. evaluated a de-identification engine used by the University of Pittsburgh Medical Center [8]. The software uses a complex set of rules like dictionaries, the Unified Medical Language System, and pattern-matching algorithms. Special attention was paid to the detection of accession numbers in pathology reports. In the first evaluation, 10.7 % of accession numbers were missed. After reprogramming of the software, only 0.7 % were missed.

Most studies were done on English reports and only a few studies, using a small sample size, developed an engine based on German grammar and vocabulary. Toepfer et al. successfully extracted information from 100 transthoracic echocardiography reports, and Bretschneider et al. evaluated 40 radiological reports [11, 12]. Ruch et al. analyzed a large set of 1000 mixed medical records with 80 784 tokens [13]. Most of them were in French, while less than 1 % was written in English and even less in German.

In the analyzed sample of pathology, medical, operative, and radiological reports, the software was able to reliably de-identify the documents for export to the scientific collaboration partners. After training, 100 % of direct and 99.3 % of indirect identifiers were detected.

Surprisingly, even misspelled words, transposed digits, and abbreviations were reliably tagged. Another interesting ability of the software was the differentiation between the use of time to specify a certain position in the breast, which must not be de-identified, versus the time of the day, which has to be de-identified.

The software faces some limitations that leave room for future improvements. One major problem in the design of the program was the concept of a “blacklist”. Once a word is added to this list, it will be de-identified in all reports. This can be seen in the specificity and PPV that do not improve after training. Particularly pro-
blematic were proper names in medical terms, e.g. “Elston and Ellis”, “Morison’s pouch”, or “Bochdalek Hernie.” Although this does not affect data security, important medical information might be lost. This problem was communicated to the developer and an editable “whitelist” (added words will not be de-identified) will be integrated in an updated version.

At the moment, medical terms which can also be a name such as “Herz” are not in the whitelist to assure that names are de-identified reliably. After sufficient training the software should be able to decide whether it is a medical term or a name dependent on the context. Therefore, medical terms and names should be annotated appropriately.

Another limitation was the missing interpretation of structured information included in the XML tags. For example, <patient>, <name>, or <address> was not used to find the patient identifier. Not only is the information lost, but the display of the XML tags decreases the readability for a human rater and therefore the sensitivity. Unfortunately, the inclusion of this information in the algorithm is complex. Ultimately, however, the file structure is different for every data set, and no universal solution may be found. The performance could surely be improved significantly if the content of the analysis is extended to the tags. Support for XML-based CDA formats is currently under development.

Varying results for direct and indirect identifiers as well as for the different types of texts might be a result of the algorithm. The results of this study will be used to further improve the software.

Despite all of these limitations, the performance of the de-identification software was more than sufficient to safely release the reports to external partners. The collaboration partners were considered reliable associates that handle the data with care on secure servers and do not try to exploit it or make it publically available. Greater safety measures must be taken if the data is to be secured from attackers with criminal intent. There-

Fig. 2 Accuracy of the de-identification software for the different kinds of reports (operative, pathology in XML and plain text, radiology and medical reports) after several steps of training. The first row shows the sensitivity for the detection of direct A and indirect B identifiers. The second row shows the specificity for the exclusion of medical terms C and filler words D from the de-identification.

fore, as shown by Gupta et al., iterative improvements and evaluations of a software algorithm must follow to optimize performance [8].

The performance of the algorithm was compared to manual segmentation that served as a reference standard. Douglass et al. showed that even human readers are not perfect [14]. First, a human evaluation is expensive and slow. Readers were able to scan about 18 000 words per hour or 90 incidents per hour. In our study, we did not record the time effort, but a rough calculation yields the same result. Scanning 5 000 000 words at the before mentioned rate of four hours a day and three days a week took us half a year. Second, even at this slow rate, the sensitivity of a human reader is far from perfect. Douglass et al. evaluated three highly motivated readers with an average sensitivity of 0.81, 0.94, and 0.98. However, the main problem is that the performance of a human is highly dependent on motivation, fatigue, and individual knowledge. These are all factors that cannot be standardized or controlled.

If neither computer nor human is perfect, how can 100 % de-identification be realized in the future to protect patients even from criminal attackers? If it is not possible to retrospectively find all relevant tokens, the only possibility is not to include them in the first place or to mark them prospectively. For example, if protected health information is only included in the header, it could be removed and only the corpus containing the medical information is exported. However, the healthcare professionals writing the report must be instructed not to include critical information in the text, especially cross-references like: date of pre-examinations, image of finding in a pre-examination, or acquisition numbers. This information is most likely to be included in the text. A human who is used to including this information in plain text is prone to errors. To erase all potential sources of error, the human input must be highly structured and running text must be omitted completely in the exported file. Due to the fact that the proposed de-identification software missed identifiers after training, a final check of the texts by an authorized person remains necessary.

The introduction of BI-RADS in the late 1980 s laid the groundwork for this kind of structured reporting [15, 16]. Breast lesions were allocated to categories ranging from benign to histologically proven malignancy. Further patient management is determined accordingly. By the year 2000, the use of BI-RADS was widespread. 93 % of surveyed radiologists reported that they always use BI-RADS and only 3 of 211 say that they never use it. With structured reporting, vague expressions or individual formulations are reduced. This makes the report more accurate and easier to read. Furthermore, findings can be included in a database for future research, running text reports can be created for the referring physician, and reports can even be created in different languages, paving the way for global teleradiology [17–20]. The BI-RADS system was adapted for liver (LI-RADS), lung (Lung-RADS) and prostate (PI-RADS) lesions, and further reporting systems are being developed [21–23]. Unfortunately, not all studies are in favor of structured reporting. Johnson et al. found a decrease in accuracy and completeness in repeated analysis of cranial magnetic resonance scans. In particular, a higher time consumption and the lack of “artistic freedom” was criticized by the radiologists [24]. The need for fast and reliable de-identification of plain text in medical reports will probably continue until precise and easy-to-use structured reporting applications are developed and become the standard in the clinical routine and science.

**Conclusion**

Working with medical reports is a balancing act between sufficiently protecting the patient’s privacy and efficiently working with the data. If external institutions are involved, reports must be de-identified reliably. Training of the proposed de-identification tool improved its accuracy. Training with roughly 100 edited reports enabled quite reliable detection and labeling of sensitive data in different types of medical reports. Due to the fact that the proposed de-identification software missed identifiers after training, a final check of the texts by an authorized person remains necessary.

**CLINICAL RELEVANCE OF THE STUDY**

- Data security has to be guaranteed for medical reports throughout collaborations with external institutions
- Semi-automated software-based de-identification can be deployed for this purpose
- The proposed software reliably provided de-identified medical reports after training with roughly 100 edited reports
- Due to the fact that the proposed de-identification software missed identifiers after training, a final check of the texts by an authorized person remains necessary

**ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMWi</td>
<td>The Federal Ministry of Economics and Technology in Germany (Bundesministerium für Wirtschaft und Energie)</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical Document Architecture</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Health Insurance Portability and Accountability Act</td>
</tr>
<tr>
<td>ML</td>
<td>Machine learning</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>Pxx</td>
<td>Performance after training with xx reports</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>UHE</td>
<td>University Hospital Erlangen</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
</tbody>
</table>

**Conflict of interest**

**Compliance with ethical standards**

This retrospective study was conducted in accordance with the guidelines of the Declaration of Helsinki and approved by the
Ethics Committee of the University Hospital Erlangen. The use of written informed consent was waived by the ethics committee.

Conflict of interest
Matthias Ihle and Andrea Grandjean are employees of the Averbis GmbH, Freiburg, Germany. They provided the software within the context of the Smart Data Program in the KDI project of the Federal Ministry for Economic Affairs and Energy, Germany (01MT14001E). They did not participate in the design, conduct and evaluation of the study. Therefore, the authors declare that no competing interests exist.

Funding
This research has been supported by the Smart Data Program in the KDI project of the Federal Ministry for Economic Affairs and Energy, Germany (01MT14001E).

References


One of the co-authors was erroneously not in the authors list. Therefore Michael Uder was added by an erratum in the online version.