Thieme

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 Seuss[1], Peter Dankerl[1], Matthias Ihle[2], Andrea Grandjean[2], Rebecca Hammon[3], Nicola Kaestle[4], Peter A. Fasching[5], Christian Maier[6], Jan Christoph[6], Martin Sedlmayr[6], Michael Uder[1], Alexander Cavallaro[1], Matthias Hammon[1]

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
Maximiliansplatz 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 Projekte bei denen verschiedene Institutionen in Kooperation miteinander stehen, erfordern einen Schutz von Patientendaten durch selektive Deidentifizierung von Wörtern oder Ausdrücken. Eine automatisierte Deidentifikations-Software wurde entwickelt und anhand verschiedener medizinischer Berichte, zuerst ohne und anschließend nach Anpassung des Algorithmus an die Textstruktur, getestet.

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

- Die softwarebasierte Deidentifikation von vertraulichen Patientendaten wird durch "Big Data" immer wichtiger
- Eine Deidentifikations-Software wurde entwickelt und im Rohzustand sowie nach manuellem Training getestet
- Nach dem Training mit etwa 100 korrigierten Texten arbeitete der Algorithmus relativ zuverlässig
- Eine abschließende Kontrolle der Texte durch eine autorisierte Person ist dennoch erforderlich

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

- Seuss H, Dankerl P, Ihle M et al. Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics. Fortschr Röntgenstr 2017; 189: 661–671

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-

▶ **Table 1** Protected health information (PHI) defined by HIPAA.

▶ **Tab. 1** Geschützte Gesundheitsdaten, definiert vom HIPAA.

Type of information	Explanatory notes
names	both full and partial, but not initials
locations	all geographic subdivisions smaller than a state, including street, address, city, county, precinct, zip code, and their equivalent geocodes
dates	all elements of dates (except years) for dates directly related to an individual, including birth date, admission date, discharge date, date of death, ages > 89 years, all elements of dates (including year) indicative of an age over 89 years. Such ages and elements may be aggregated into a single category of age 90 or older
telephone numbers	
fax numbers	
electronic mail addresses	
social security numbers	
medical record numbers	
health plan beneficiary numbers	
account numbers	
certificate/license numbers	
vehicle identifiers	includes vehicle serial numbers and license plate numbers
device identifiers and serial numbers	not restricted to medical devices
Web Universal Resource Locators (URLs)	
Internet Protocol (IP) address numbers	
biometric identifiers	includes finger and voice prints
any other unique identifying number, code, or characteristic e. g., full photographic images of full faces, scars or tattoos	
(and any comparable images)	

ted at all times. Furthermore, identifiers of the hospital and its employees must also be extracted. Data that must not be secured include, for example, medical content and negations.

The basis of 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 4671 pathology reports, 2804 medical reports, 1008 operative reports and 6223 radiology reports. The pathology reports came from the Institute of Pathology of the University Hospital Erlangen (UHE) and existed in two different formats, 4105 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.

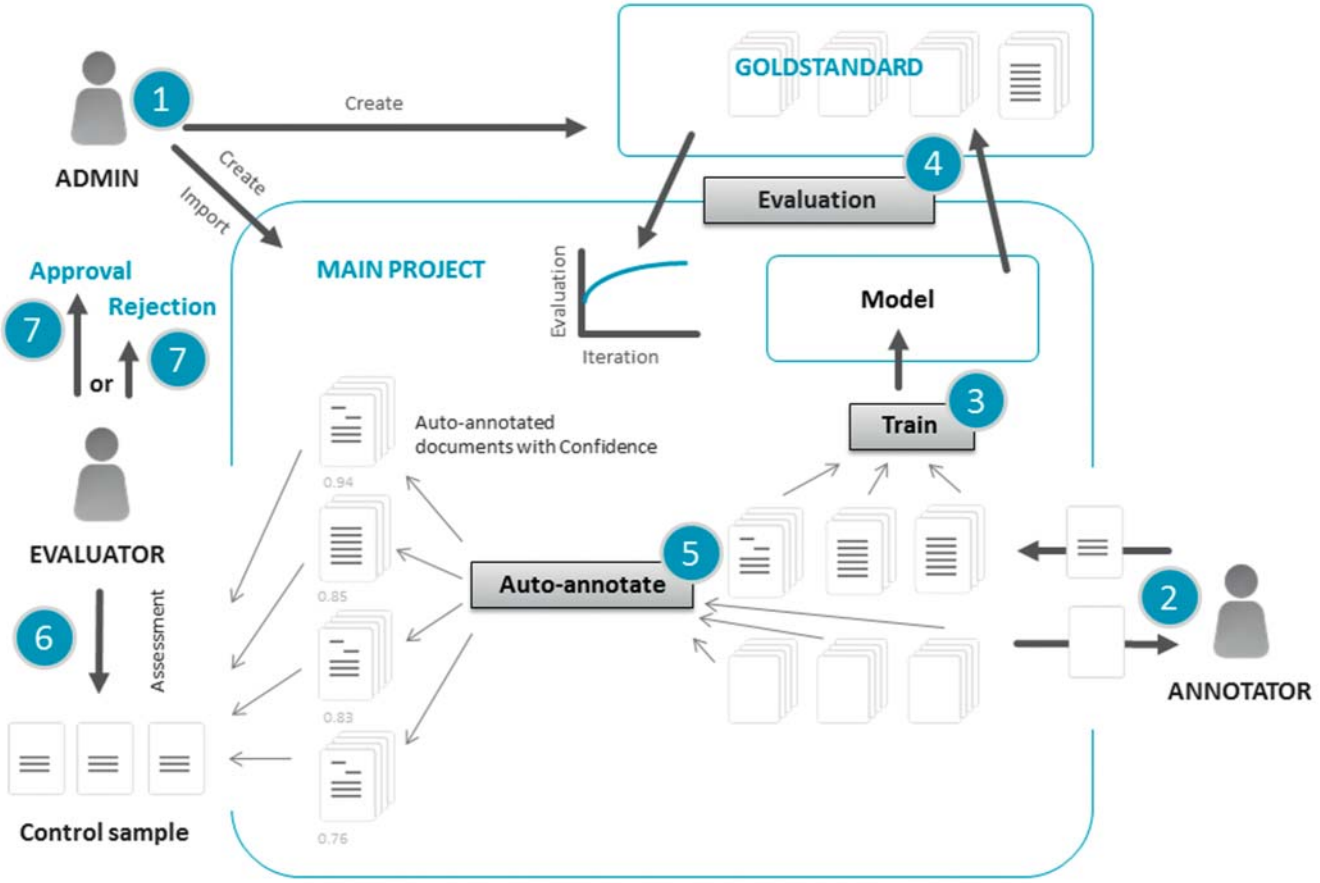
Re-training the model

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 the 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



▶ **Fig. 1** The de-identification process: 1. The administrator creates the project and imports the reports for de-identification. 2. The annotator labels the findings. 3. After annotation of a set of reports, the model can be retrained and 4. The results are compared to the gold standard. 5. Now all documents can be automatically annotated and the algorithm calculates its confidence in its accuracy. 6. A human reviewer checks sample reports and can 7. Approve or reject the annotation.

▶ **Abb. 1** Der Deidentifikationsprozess: 1. Der Studienverantwortliche erstellt das Projekt und importiert die Berichte für die Deidentifikation. 2. Die Funde werden durch den Annotator markiert. 3. Nachdem der erste Satz an Berichten annotiert wurde, wird das Modell an diesem trainiert und 4. dessen Ergebnisse mit dem Goldstandard verglichen. 5. Nun werden alle weiteren Dokumente automatisch annotiert und das Programm

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 14 706 manually corrected reports, a total of 1400 documents with 5000 921 words were analyzed for this study; of these, 4563 words were direct and 18 976 indirect identifiers; 93 199 words were of medical importance; and 4884 183 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.

18 286 (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.

▶ **Tab. 2** Ergebnisse der Deidentifikationsleistung des Programms. Für direkte und indirekte Identifikatoren sind die richtig Positiven, falsch Negativen, Gesamtzahl der Funde und die Sensitivität dargestellt, aufgeschlüsselt nach den verschiedenen Berichten. Für medizinischen Begriffe und Füllwörter sind die richtig Negativen, die falsch Positiven, die Gesamtzahl der Funde und die Spezifität dargestellt.

		direct identifier				indirect identifier				medical terms				filler words			
		TP	FN	total	sensitivity	TP	FN	total	sensitivity	TN	FP	total	specificity	TN	FP	total	specificity
1008 operation reports (n = 250)	0	202	61	263	76.8 %	89	54	143	62.2 %	2564	94	2658	96.5 %	22 578	88	22 666	99.6 %
	25	230	0	230	100.0 %	133	11	144	92.4 %	2417	99	2516	96.1 %	20 086	68	20 154	99.7 %
	50	480	6	486	98.8 %	306	6	312	98.1 %	4887	144	5031	97.1 %	43 062	126	43 188	99.7 %
	100	354	1	355	99.7 %	228	6	234	97.4 %	4299	171	4470	96.2 %	37 935	105	38 040	99.7 %
	250	359	1	360	99.7 %	262	3	265	98.9 %	4017	157	4174	96.2 %	45 904	129	46 033	99.7 %
	500	323	0	323	100.0 %	240	2	242	99.2 %	4595	200	4795	95.8 %	36 455	89	36 544	99.8 %
566 pathology reports (XML) (n = 250)	0	12	60	72	16.7 %	470	120	590	79.7 %	176	19	195	90.3 %	29 295	24	29 319	99.9 %
	25	59	21	80	73.8 %	423	10	433	97.7 %	222	16	238	93.3 %	23 380	17	23 397	99.9 %
	50	166	13	179	92.7 %	1099	23	1122	98.0 %	342	23	365	93.7 %	54 531	81	54 612	99.9 %
	100	158	2	160	98.8 %	1151	21	1172	98.2 %	408	32	440	92.7 %	58 285	81	58 366	99.9 %
	250	172	1	173	99.4 %	838	4	842	99.5 %	353	41	394	89.6 %	46 777	113	46 890	99.8 %
	500	138	1	139	99.3 %	947	5	952	99.5 %	389	29	418	93.1 %	58 437	128	58 565	99.8 %
4105 pathology reports (n = 300)	0	16	0	16	100.0 %	16	23	39	41.0 %	3953	208	4161	95.0 %	35 545	37	35 582	99.9 %
	25	58	0	58	100.0 %	50	11	61	82.0 %	4030	105	4135	97.5 %	35 755	27	35 782	99.9 %
	50	101	0	101	100.0 %	98	20	118	83.1 %	8506	198	8704	97.7 %	84 170	82	84 252	99.9 %
	100	87	0	87	100.0 %	97	18	115	84.3 %	6915	203	7118	97.1 %	74 359	59	74 418	99.9 %
	250	65	0	65	100.0 %	75	6	81	92.6 %	9467	230	9697	97.6 %	96 795	93	96 888	99.9 %
	500	67	0	67	100.0 %	82	7	89	92.1 %	7807	219	8026	97.3 %	76 458	83	76 541	99.9 %
	1000	57	0	57	100.0 %	87	6	93	93.5 %	9656	179	9835	98.2 %	90 406	87	90 493	99.9 %
6223 radiology reports (n = 300)	0	13	0	13	100.0 %	101	2	103	98.1 %	394	38	432	91.2 %	3398	15	3413	99.6 %
	25	15	1	16	93.8 %	125	3	128	97.7 %	459	35	494	92.9 %	3956	15	3971	99.6 %
	50	27	1	28	96.4 %	237	5	242	97.9 %	827	64	891	92.8 %	7119	11	7130	99.8 %
	100	34	0	34	100.0 %	281	8	289	97.2 %	980	43	1023	95.8 %	8453	8	8461	99.9 %
	250	32	0	32	100.0 %	262	4	266	98.5 %	1028	53	1081	95.1 %	8900	21	8921	99.8 %
	500	34	0	34	100.0 %	234	0	234	100.0 %	943	74	1017	92.7 %	8152	12	8164	99.9 %
	1000	37	0	37	100.0 %	289	1	290	99.7 %	1089	57	1146	95.0 %	9309	24	9333	99.7 %

▶ **Table 2** (Continuation)

		direct identifier				indirect identifier				medical terms				filler words			
		TP	FN	total	sensitivity	TP	FN	total	sensitivity	TN	FP	total	specificity	TN	FP	total	specificity
2804 medical reports (n = 300)	0	45	61	106	42.5 %	683	124	807	84.6 %	853	89	942	90.6 %	356 273	1 88	356 461	99.9 %
	25	73	19	92	79.3 %	746	73	819	91.1 %	833	81	914	91.1 %	348 946	2271	351 217	99.4 %
	50	153	12	165	92.7 %	2060	29	2089	98.6 %	1391	174	1565	88.9 %	633 309	3939	637 248	99.4 %
	100	186	1	187	99.5 %	1518	43	1561	97.2 %	1675	146	1821	92.0 %	672 194	291	672 485	100.0 %
	250	175	1	176	99.4 %	1550	15	1565	99.0 %	1120	128	1248	89.7 %	662 805	398	663 203	99.9 %
	500	184	1	185	99.5 %	1517	17	1534	98.9 %	1385	160	1545	89.6 %	534 414	383	534 797	99.9 %
	1000	187	0	187	100.0 %	1992	10	2002	99.5 %	1556	154	1710	91.0 %	647 385	264	647 649	100.0 %
all 14 706 reports (n = 1400)	0	288	182	470	61.3 %	1359	323	1682	80.8 %	7940	448	8388	94.7 %	447 089	352	447 441	99.9 %
	25	435	41	476	91.4 %	1477	108	1585	93.2 %	7961	336	8297	96.0 %	432 123	2398	434 521	99.4 %
	50	927	32	959	96.7 %	3800	83	3883	97.9 %	15 953	603	16 556	96.4 %	822 191	4239	826 430	99.5 %
	100	819	4	823	99.5 %	3275	96	3371	97.2 %	14 277	595	14 872	96.0 %	851 226	544	851 770	99.9 %
	250	803	3	806	99.6 %	2987	32	3019	98.9 %	15 985	609	16 594	96.3 %	861 181	754	861 935	99.9 %
	500	746	2	748	99.7 %	3020	31	3051	99.0 %	15 119	682	15 801	95.7 %	713 916	695	714 611	99.9 %
	1000	281	0	281	100.0 %	2368	17	2385	99.3 %	12 301	390	12 691	96.9 %	747 100	375	747 475	99.9 %

▶ **Table 3** The 25 most common false-positive findings of medical terms in German and their English translation.

▶ **Tab. 3** Die 25 von dem Programm am häufigsten als falsch positiv markierten medizinischen Begriffe in Deutsch und deren englische Übersetzung.

medical term	translation	tags
Herz	heart	99
Winkel	angle	81
ED MaCa	FD BrCA	59
Leber	liver	25
Milz	spleen	24
Oberes	upper	23
Port	port	22
Brust	breast	20
Gabe von Blut	administration of blood	19
Springer	circulator	19
Tasche	pocket	16
Weite	width	15
Hals	throat	14
Spitze	tip	13
Unterfeld	lower lung field	13
Herd	lesion	12
Kalk	calcification	11
Hals	throat	11
Skinsparing	Skinsparing	10
Becken	pelvis	10
Höhe	height	8
Kleine	small	7
LO VO	LO VO	7
Tasche	pocket	7
VU Bild	pre-examination image	7

(P250), 73.2 % (P500), 77.6 % (P1000), and the negative predictive value 99.9 % (P0), 100 % (P25 – P1000).

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.

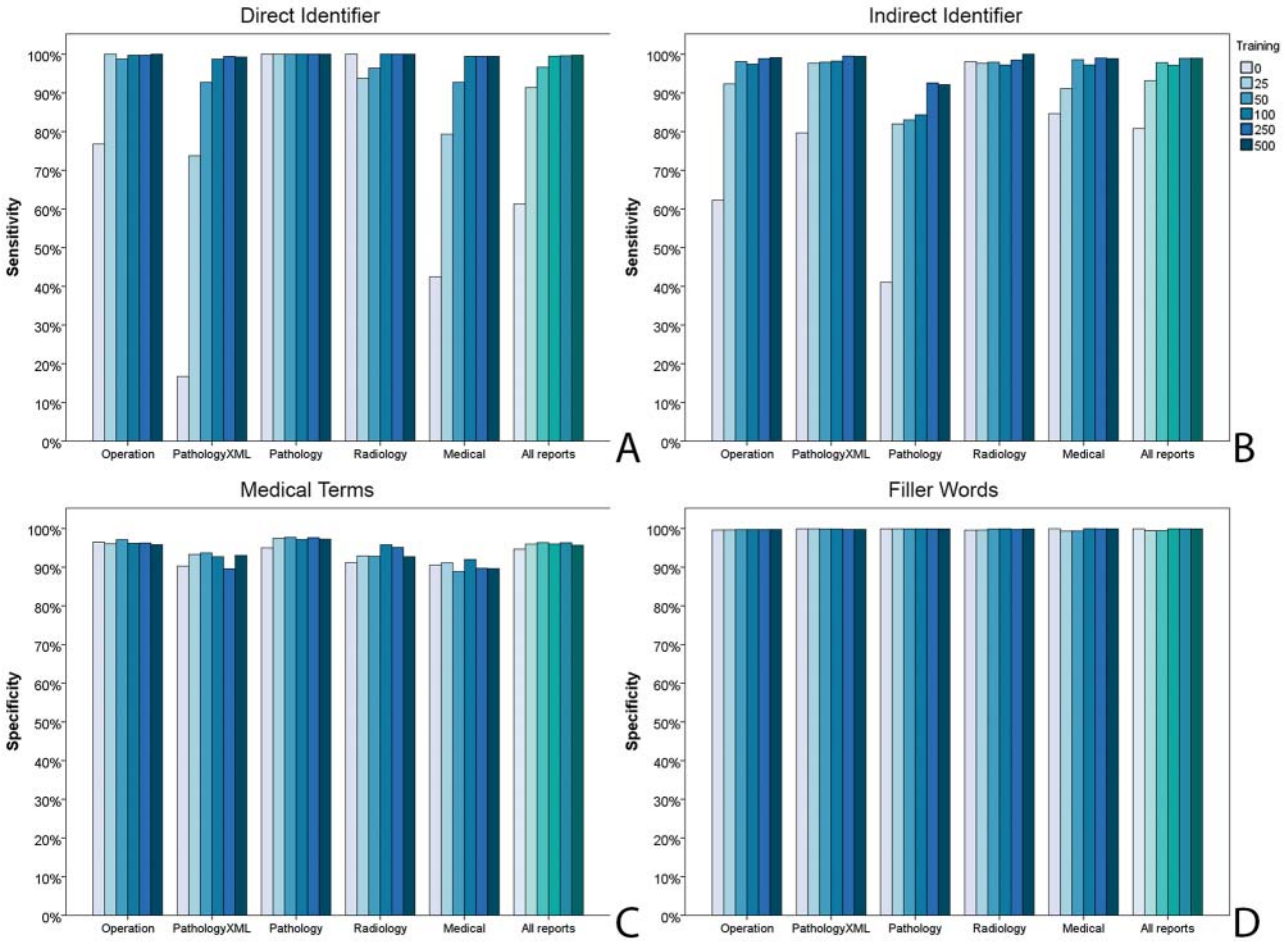
Gupta et al. evaluated a de-identification engine used by the University of Pittsburgh Medical Center [8]. The software uses a complex set of rules like dictionaries, the Unified Medical Language System, and pattern-matching algorithms. Special attention was paid to the detection of accession numbers 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-



▶ **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.

▶ **Abb. 2** Annotationsleistung der Deidentifikationssoftware für die unterschiedlichen Berichte (Operationsbericht, pathologischer Bericht im XML-Format und als Fließtext, radiologische Befunde und Arztbriefe) nach den jeweiligen Trainingsstufen. In der ersten Reihe ist die Sensitivität für die Detektion von direkten **A** und indirekten **B** Identifikatoren dargestellt. Die zweite Reihe bildet die Spezifität für das Nicht-Deidentifizieren von medizinschem Inhalt **C** und von Füllwörtern **D** ab.

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-

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 5000 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 to not 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

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

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 (01MT14 001E). They did not participate in the design, conduction 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 (01MT14 001E).

References

[1] Edelstein L. The Hippocratic oath, text, translation and interpretation. Baltimore: The Johns Hopkins press; 1943

[2] Nass SJ, Levit LA, Gostin LO. Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule. Washington (DC): National Academies Press (US). 2009

[3] U.S. Department of Health & Human Services. Federal Policy for the Protection of Human Subjects ("Common Rule"). 1991

[4] European Parliament, Council of the European Union. Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). 2002

[5] Neamatullah I, Douglass MM, Lehman LW et al. Automated de-identification of free-text medical records. BMC medical informatics and decision making 2008; 8: 32

[6] Kushida CA, Nichols DA, Jadrnicek R et al. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Medical care 2012; 50: S82 – S101

[7] Sweeney L. Computational disclosure control: a primer on data privacy protection. PhD thesis, Massachusetts Institute of Technology 2001 http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/sweeney-thesis-draft.pdf. Accessed February 22, 2017

[8] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. American journal of clinical pathology 2004; 121: 176 – 186

[9] Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning. 2001

[10] Thomas SM, Mamlin B, Schadow G et al. A successful technique for removing names in pathology reports using an augmented search and replace method. Proceedings / AMIA Annual Symposium AMIA Symposium. 2002: 777 – 781

[11] Toepfer M, Corovic H, Fette G et al. Fine-grained information extraction from German transthoracic echocardiography reports. BMC medical informatics and decision making 2015; 15: 91

[12] Bretschneider C, Zillner S, Hammon M. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. Proceedings of BioNLP 2013: 27 – 35

[13] Ruch P, Baud RH, Rassinoux AM et al. Medical document anonymization with a semantic lexicon. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000: 729 – 733

[14] Douglass M, Clifford GD, Reisner A et al. Computer-assisted de-identification of free text in the MIMIC II database. Computers in Cardiology 2004; 2004: 341 – 344

[15] Burnside ES, Sickles EA, Bassett LW et al. The ACR BI-RADS experience: learning from history. J Am Coll Radiol 2009; 6: 851 – 860

[16] Margolies LR, Pandey G, Horowitz ER et al. Breast Imaging in the Era of Big Data: Structured Reporting and Data Mining. Am J Roentgenol American journal of roentgenology 2016; 206: 259 – 264

[17] Hobby JL, Tom BD, Todd C et al. Communication of doubt and certainty in radiological reports. Br J Radiol 2000; 73: 999 – 1001

[18] Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image. Br J Radiol 1997; 70: 1085 – 1098

[19] Hawkins CM, Hall S, Zhang B et al. Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports. J Digit Imaging 2014; 27: 581 – 587

[20] Durack JC. The value proposition of structured reporting in interventional radiology. Am J Roentgenol American journal of roentgenology 2014; 203: 734 – 738

[21] Mitchell DG, Bruix J, Sherman M et al. LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology (Baltimore, Md) 2015; 61: 1056 – 1065

[22] American College of Radiology. Lung CT Screening Reporting and Data System (Lung-RADS). https://www.acr.org/Quality-Safety/Resources/LungRADS. Accessed February 22, 2017

[23] Barentsz JO, Richenberg J, Clements R et al. ESUR prostate MR guidelines 2012. Eur Radiol 2012; 22: 746 – 757

[24] Johnson AJ, Chen MY, Swan JS et al. Cohort study of structured reporting compared with conventional dictation. Radiology 2009; 253: 74 – 80

Erratum to the article "Seuss H, Dankerl P, Ihle M et al. Semi-automated De-identification of German Content Sensitive Reports for Big Data Analytics. Fortschr Röntgenstr 2017; 189: DOI 10.1055/s-0043-102939"

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