The need for harmonized structured documentation and chances of secondary use – Results of a systematic analysis with automated form comparison for prostate and breast cancer

Rainer Krumm a,*, Axel Semjonow b, Joke Tio c, Heike Duhme d, Thomas Bürkle e, Jörg Haier d, Martin Dugas a,1, Bernhard Breil a,1

a Institute of Medical Informatics, University of Münster, Albert-Schweitzer-Campus 1, A11, D-48149 Münster, Germany
b University Hospital Münster, Department of Urology – Prostate Center, Albert-Schweitzer-Campus 1, A1, D-48149 Münster, Germany
c University Hospital Münster, Department of Obstetrics and Gynecology, Albert-Schweitzer-Campus 1, A1, D-48149 Münster, Germany
d University Hospital Münster, Comprehensive Cancer Center Münster, Waldeyerstr. 1, D-48149 Münster, Germany
e University Erlangen-Nürnberg, Chair of Medical Informatics, Krankenhausstraße 12, D-91054 Erlangen, Germany

Article history:
Received 16 August 2013
Accepted 7 April 2014
Available online 18 April 2014

Keywords:
Semantic interoperability
Automated form comparison
Common data elements
Secondary use
Standardization
Clinical documentation

Abstract

Introduction: Medical documentation is a time-consuming task and there is a growing number of documentation requirements. In order to improve documentation, harmonization and standardization based on existing forms and medical concepts are needed. Systematic analysis of forms can contribute to standardization building upon new methods for automated comparison of forms. Objectives of this research are quantification and comparison of data elements for breast and prostate cancer to discover similarities, differences and reuse potential between documentation sets. In addition, common data elements for each entity should be identified by automated comparison of forms.

Materials and methods: A collection of 57 forms regarding prostate and breast cancer from quality management, registries, clinical documentation of two university hospitals (Erlangen, Münster), research datasets, certification requirements and trial documentation were transformed into the Operational Data Model (ODM). These ODM-files were semantically enriched with concept codes and analyzed with the compareODM algorithm. Comparison results were aggregated and lists of common concepts were generated. Grid images, dendrograms and spider charts were used for illustration.

Results: Overall, 1008 data elements for prostate cancer and 1232 data elements for breast cancer were analyzed. Average routine documentation consists of 390 data elements per disease entity and site. Comparisons of forms identified up to 20 comparable data elements in cancer conference forms from both hospitals. Urology forms contain up to 53 comparable data elements with quality management and up to 21 with registry forms. Urology documentation of both hospitals contains up to 34 comparable items with international common data elements. Clinical documentation sets share up to 24 comparable data elements with trial documentation. Within clinical documentation administrative items are most common comparable items. Selected common medical concepts are contained in up to 16 forms.

Discussion: The amount of documentation for cancer patients is enormous. There is an urgent need for standardized structured single source documentation. Semantic annotation is time-consuming, but enables automated comparison between different form types, hospital sites and even languages. This approach can help to identify common data elements in medical documentation. Standardization of forms and building up forms on the basis of coding systems is desirable. Several comparable data elements within the analyzed forms demonstrate the harmonization potential, which would enable better data reuse.

Conclusion: Identifying common data elements in medical forms from different settings with systematic and automated form comparison is feasible.

© 2014 Elsevier Inc. All rights reserved.

* Corresponding author.
E-mail addresses: Rainer.Krumm@uni-muenster.de (R. Krumm), Axel.Semjonow@ukmuenster.de (A. Semjonow), Joke.Tio@ukmuenster.de (J. Tio), Heike.Duhme@ukmuenster.de (H. Duhme), Thomas.Buerkle@uni.med.uni-erlangen.de (T. Bürkle), Joerg.Haier@ukmuenster.de (J. Haier), Dugas@uni-muenster.de (M. Dugas), Bernhard.Breil@uni-muenster.de (B. Breil).
1 Authors contributed equally.
1. Introduction

Medical documentation is a required, but time-consuming task and therefore in the focus of health informatics research. The official journal of the German Medical Association titled 2009 one article “Documentation in medicine – it is a madness” [1] and addressed the problem of documentation reality. There is high documentation workload and a huge variety of forms for different purposes with redundant data capture.

The documentation process is very complex and sophisticated especially in oncology. Nevertheless, there is a lack of standardization which results in a huge variety of different forms in hospitals even for the same disease areas. This work is focused on the analysis of documentation items of widespread cancer diseases. Prostate cancer is the most common cancer (other than skin cancer) among American men and breast cancer the second leading cause of cancer-related death in women [2,3]. Both cancer documentation sets consist of many forms from different sources and contain also redundant items. It was reported previously that e.g. for breast cancer up to 65% of data elements are gathered redundantly [4]. An optimization of this documentation process is desirable because documentation should burden the physician as few as possible [5].

This becomes even more relevant as physicians spend more than 25% of their daily working time for documentation tasks and hereby almost as much time as for direct patient care [6,7]. These numbers only consider routine documentation within a hospital. In addition, parallel and redundant documentation for quality management, registries and clinical research increase the physicians’ workload.

The documentation for these different purposes is in large part done within different systems and structures. “The existence of parallel independent documentation systems leads to a tremendous workload, and thus hinders acceptance among physicians” [8]. Patient care forms usually consist of many free-text elements, whereas research and registry-forms are highly structured.

Studies show that there is a re-use potential for these purposes if the original information is documented in a structured way [9]. Prokosch reviewed the potential of reusing the electronic health record (EHR) [10] and Kush introduced documentation according to the single-source concept to reduce the physician's workload [11]. Further projects in this context demonstrate the high potential of this research field [12].

Without secondary use and structured documentation a lot of resources are needed to fulfill the documentation requirements for research and registries. This documentation is up until today usually performed by manual review of clinical free-text. This is an inefficient and error-prone process because the same information needs to be documented two or three times (e.g. writing of physician’s letter and extracting information from it to complete quality management forms). A similar aspect can be observed in the context of clinical studies, as Getz reported that the “protocol designs are becoming more demanding and burdensome on investigative site personnel”. Between 1999 and 2005 the workload for research increased by 10.5% per year [13].

At the moment many studies are dealing with optimization of electronic medical records (EMR) or electronic health record (EHR) systems but there is still a lack of documentation standardization which leads to heterogeneity in data representation [14].

Future work has to focus not only on the improvement of EHR systems but on the standardization of medical forms and their data items and to provide documentation standards that meet the different purposes by a single source documentation system. This need is emphasized by Ries et al., who analyzed oncological documentation in Germany and concluded that “none of the existing German cancer datasets (e.g. ADT or GEKID) meets clinical documentation reality” [8].

Therefore, a systematic analysis is needed to identify common concepts and elements based upon documentation reality. To identify redundant documentation items it is necessary to compare and harmonize medical data items. Semantic enriched data items can enable automated comparison. Systematized Nomenclature of Medicine (SNOMED) [15] or Unified Medical Language Systems (UMLS) [16] are very important in this context [17].

This work applies and extends compareODM, a recently published method to compare coded forms [18]. In contrast to the previous paper, this research goes beyond comparison of two forms and addresses a systematic analysis of the documentation landscape from two major disease entities at two university hospital sites including routine documentation, quality management, certification, research and cancer registration on the basis of real forms. Therefore, new methods to compare and visualize groups of forms were applied, for example spider charts. It was required to extend the compareODM method to conduct these analyses: the comparison of code lists (value lists) was integrated and the output of compareODM was enhanced for statistical analysis. With the compareODM Paper [18] the feasibility to compare two given ODM-forms was demonstrated. In the current research common data elements in tumor documentation for two chosen diseases were analyzed and identified.

To our knowledge, an analysis of such amount of forms from clinical documentation is not available in the literature. Results of this paper are precise lists of data elements and can contribute to improved information systems.

The scope of this work is an assessment of standardization opportunities based on a systematic and automated comparison of oncological forms in the context of prostate and breast cancer. It is well known that standardization can lead to better interoperability [19].

2. Objectives

The overall objective of this work is to analyze the current documentation landscape for two cancer entities, breast and prostate cancer. Specifically, we want to address the following research questions:

1. Is it possible to quantify the amount of forms and data elements?
2. Is there a chance of reusing elements for secondary use purposes?
3. Which similarities exist between documentation for quality management, registries, research and clinical routine?
4. What are the differences between clinical routine documentation in two German university hospitals?
5. What are the top 30 common data elements for each disease entity?

3. Methods and materials

3.1. Form analysis process

3.1.1. Form collection and coding of data elements

The documentation landscape of two common cancer entities (breast and prostate cancer) was analyzed to determine currently used data elements. For both diseases there exist structured forms. Medical forms for these cancers entities were collected from two university hospitals (Erlangen and Münster), e.g. forms for medical history, forms regarding surgery and different types of...
datasets (e.g. international common data elements, clinical trial management and certification as well as different research documents) were available for analysis. In addition, a web search for available forms (for instance out of date or only administrative items) 57 forms were collected and after excluding forms (like technical implementation or usability issues) 86 forms were collected and after excluding forms without matching form for comparison 86 forms were collected and after excluding forms containing only one score 86 forms were collected and after excluding forms without permission 86 forms were collected and after excluding forms without permission.

Fig. 1. Number of forms in the analysis process. 86 forms were collected and after excluding forms (for instance out of date or only administrative items) 57 forms were available for analysis.

3.2. Documentation sources

Documentation sources are characterized and classified according to the process in which they are used. Empty documentation forms were collected from sources of the following four categories in order to cover all relevant documentation elements. At this point only a short overview is presented. A detailed description and characterization of individual forms can be found within the Supplement (Supplement file 2) as well as a full list of material (Supplement file 1). Afterwards the forms are referenced with their file names. The file names are mentioned in italic letters without file extension (".xml").

- Routine documentation forms from two university hospitals (Münster, Erlangen).
- Breast cancer (14 forms).
- Prostate cancer (17 forms).
- Registries.
  - Epidemiologic Cancer Registry North-Rhine Westphalia (Epi-NRW) [25].
  - Dataset of the Society of Epidemiological Cancer Registries in Germany (GEKID) [26].
  - Finish Cancer Registry (Finish_register) [27].
  - Certification & quality management documents: 20 forms.
  - ADT¹ – Basic Cancer Documentation [29].
  - ONDIS: oncoligic aftercare, documentation and information system [30].
  - prostate cancer center certification by German Cancer Society (Onkozert prostate).
  - Quality assurance program for breast cancer centers in the State of North Rhine-Westphalia (Certification_breast).
  - AQUA²: Statutory Mandatory Quality Management: questionnaire for breast surgery (AQUA_breast) [32].
  - Research documentation.
    - Dataset: Deutsches Prostatakarzinom Konsortium e.V. (DPKK) [33], a cross-institutional research network on prostate cancer (DPKK_prostate).
    - Dataset: Common data elements from CPCTR (Cooperative Prostate Cancer Tissue Resource) (CDE_prostate) [34].
    - Trial documentation: ALTTO (Adjuvant Lapatinib And/Or Trastuzumab Treatment Optimization) (NCT00490139) [35].

All collected forms were transformed into the CDISC Operational Data Model (ODM) [20], which is a standardized XML-format. For this transformation all data items from those forms were extracted manually and described by a data type (e.g. boolean, string), an item name and a concept code. These manually created element-lists were converted into the final ODM-files with the ODMConverter software (available on http://cran.r-project.org [21]). All data items of each form had to be coded with terminology codes to enable automated comparison. Automated comparison requires structured and coded forms and can be performed by a computer in contrast to elaborate manual comparison by humans. In the following “automated form comparison” refers to form comparison by the presented algorithm. Coding was performed by a medical expert (RK).

Data items were coded with several existing coding systems, in particular UMLS and SNOMED CT. All forms were uploaded into the portal of medical data models (http://www.medical-data-models.org). The list of form URLs is available in the Supplement (Supplement file 1).

NCI Metathesaurus [22] was used to identify suitable codes. The best-fitting medical concept was identified and the semantic type (procedure, finding, body part, etc.) was also taken into account. If an item could not be represented by a single precoordinated code, different codes were assigned using postcoordination to represent the concept (For example: “date of diagnosis tumor finding” → tumor finding (C1274082) and date of diagnosis (C2316983)). If the information and definition of a code were not sufficient within NCI Metathesaurus additional information about SNOMED Codes within CliniciClue® Xplore [23] were used, for example: two different SNOMED Codes linked to one UMLS Code: zoledronic acid (C0257685) linked to: zoledronic acid PT 134600006 – type = pharmaceutical/biological product and to: zoledronic acid PT 395926009–type = substance).

All ODM-forms were loaded into a database. Statistical analyses were performed with SPSS [24] to determine the number of forms and items.
3.2.1. Summary of analyzed forms with documentation timeline

All forms were arranged in a timeline (Fig. 2) to provide an overview of the context in which the included forms are used. Documents were mapped to general workflow steps (medical history, diagnostic, therapy decision, therapy, aftercare). Documents that are not bound to a specific point in time were listed separately. The mapping to the workflow is also available with full details in the Supplement (Supplement file 1).

3.3. Form comparison

The coded ODM-forms were compared using compareODM [18]. A basic version for comparison of two ODM-files is available on http://cran.r-project.org [36]. Within this analysis an extended version of compareODM was implemented that is able to work with a set of forms. We added the comparison of code lists (value lists) and prepared the output for statistical analyses.

The extended R-source code is available in the Supplement (Supplement file 14). Within compareODM “two items are called identical if item name, concept code and value domain are the same. Two items are called matching if item names are different but concept codes and value domain are the same. Two items are called similar if their concept codes are the same but value domains are different.” [18]. Forms were compared to determine common data elements between different health care facilities and to examine the overlap with registry and quality management documents.

Within the scope of this work the focus is set on comparable items. These are defined as the aggregation of identical, matching and similar items. This aggregation was performed due to the complexity of prostate cancer and breast cancer form sets and especially for visualization and determination of common medical concepts needed. To determine coverage, the number of different concepts within the list of comparable items was divided by the lower number of items of involved forms.

The numbers of comparable data elements between different forms and relations of forms are illustrated with different types of visualization. To extend the compareODM approach with dendrograms and grid images and to manage the complexity of form comparisons, spider charts for the illustration of the relations within groups of forms were added. With this illustration two groups of forms are compared: One group is plotted clockwise around and the other forms are each plotted with a curve of different color. Curves represent the number of comparable items regarding two corresponding forms. Spider charts were created with Microsoft Excel®.

Form analysis was performed separately for each entity in the following steps:

1. Comparison of routine documentation of both hospitals.
2. Comparison of routine documentation of both hospitals with certification & quality management forms.
3. Comparison of routine documentation with different registry forms.
4. Comparison of routine documentation with research documentation.

Focus was set to pairwise comparison of forms with high number of comparable items and high relevance from a medical point.

Fig. 2. Arrangement of all forms in a timeline. Clinical documents and related quality management forms are mapped to a general timeline. Different phases and documentation tasks within a typical clinical workflow are presented. The four groups below the timeline have no chronological order.
of view. Different documentation standards (CDE prostate, registries and clinical trial documentation) were analyzed and set in relation to each other.

3.3.1. Lists of comparable items and common data elements
Identification of common data elements for breast and prostate cancer was done by semantic enrichment (coding of data items) of collected forms, descriptive analyses (most frequent elements) as well as systematic and automated form comparisons.

Automated form comparison was performed with an extended version of compareODM and results in matrices with all concepts that are identical, matching or similar within the comparisons. This output was analyzed with SPSS [24] and Microsoft Excel to determine the top 30 common concepts for prostate and breast cancer. Results were presented in tabular format with absolute and relative frequencies as well as associated UMLS concepts.

4. Results
57 out of 86 forms were included into the documentation analysis for prostate and breast cancer. In total, these 57 forms consist of 3565 items, which means on average 63 items per form (standard deviation 47, minimum 20, maximum 314). Included forms contain 18 general documents for cancer documentation, like registries, with total 1325 items and thereof 1300 coded items. Regarding this group of forms: the ADT base dataset consists of 449 and 447 coded items in total, the ONDIS dataset consists of 707 and 684 coded items in total and the included registries contain up to 65 items.

Routine documentation for breast and prostate cancer contains on average 390 data elements per disease entity and site. The highest number of items within routine documentation is 108 items and for certification and quality management 175 items for one form. The included selection of trial documentation contains 314 items and thereof 201 coded items.

4.1. Form comparison
Form comparison was done pairwise. Fig. 3 shows the frequencies of comparable items for typical examples. Due to the complexity only the first rows of the table are listed. The complete tables with all forms concerning breast and prostate cancer as well as general cancer forms are available in the Supplement (Supplement file 3).

Each cell represents the number of comparable items between the two related forms separated into identical items (concept, name and data type/value list identical), matching items (only name different) and similar items (concept identical, data type or value list different).

4.1.1. Prostate cancer documentation
This analysis includes 21 forms related to prostate cancer with 1008 items and thereof 1001 coded items. Routine documentation for this entity consists of 803 items, with 797 coded items. The documentation set in Münster contains 224 items and 221 coded items and in Erlangen 579 items and 576 coded items.

Comparing urology routine documentation of the two sites there are high numbers of comparable items (identical 6, matching 6, and similar 8) for the cancer conference forms. In addition, the pathology forms for biopsies share comparable items (identical 2, matching 5, similar 3). Full statistical results are provided in the Supplement (Supplement file 3) as well as graphical results of comparing documentation in Erlangen and Münster (Supplement file 4, grid images figure I.1–8 and dendrograms figure II.1–4). All in all, the documentation sets from Münster and Erlangen share 56 concepts, which corresponds to a relative overlap of 25%.

Comparing urology routine documentation forms with certification and quality management forms (e.g. ADT, ONDIS) results in up to 53 comparable items. The grid-image in Fig. 4 highlights matching items within this form set. Further graphical results can be found within the Supplement (Supplement file 4, grid-images figure III.1–7 and dendrograms figure IV.1–4).

Comparing ADT dataset with forms from Erlangen (box A in Fig. 4) shows that there are up to 11 matching and 29 similar items. Comparing urology forms from Münster with ADT results in maximum 7 matching and 7 similar items. The documentation set of Erlangen shares 161 comparable items with the ADT dataset. This demonstrates that 34% of the concepts are covered by routine documentation. In Münster there are 39 comparable items, which corresponds to a coverage of 8%.

Between routine urology forms and ONDIS (box C in Fig. 4) in Erlangen up to 17 matching and 16 similar items were identified in pairwise comparison and in Münster up to 5 matching and 6 similar items. For example Uro-E_medical-history has 8 identical items, 18 matching and 10 similar items compared to the ONDIS_diagnosis from 70 items in total.

Comparing urology routine documentation with cancer registry forms is marked with box D in Fig. 4 and illustrated with more details in Fig. 5. There are up to 21 comparable data elements between routine documentation forms and registries. The documentation set of Münster shares 10 comparable items with EpiNRW, which means that 15% of EpiNRW is covered by routine documentation. Erlangen shares 31 comparable items, which means that 48% are covered.

The comparison with common data elements (CDE_prostate) is marked with box B in Fig. 4 and the number of comparable data elements is visualized in Fig. 5. All in all, there are in one form up to 14 comparable items between routine documentation forms and CDE_prostate. The documentation set from Erlangen shares 34 concepts (36%) with CDE_prostate and the one from Münster 18 (19%).

Further results of comparing urology routine documentation with registries and research documentation can be found within the Supplement (Supplement file 5).

4.1.2. Breast cancer documentation
The analysis includes 18 forms related to breast cancer with 1232 items and thereof 1115 coded items. Routine documentation for this entity consists of 786 items, with 783 coded items. The documentation set in Münster contains 573 items and 571 coded items and in Erlangen 213 items and 212 coded items.

Comparisons in the field of gynecology were done in the same way as in the field of urology: Routine documentation of the two sites was compared. The highest number of similar items was found regarding cancer conference protocols (up to 9 similar items, box E in Fig. 6). The documentation sets from Münster and Erlangen share 58 concepts, which corresponds to a relative overlap of 27%.

Comparing gynecology routine documentation with certification and quality management, the documentation set of Münster shares 53 comparable items with the ADT data set, which shows that 11% are covered by routine documentation. In Erlangen routine documentation shares a similar number of 49 comparable items (10%). Further graphical results are available in the Supplement (Supplement file 6, grid images figure III.1–8 and dendrograms figure IV.1–4 and V.1). Detailed statistical results are provided in the Supplement (Supplement file 3).

Comparing gynecology routine documentation with registries, the documentation set from Münster shares 14 comparable items with EpiNRW, corresponding to 22% coverage by routine documentation. In Erlangen there are 25 comparable items (38%). The spider chart
Fig. 3. Output of the compareODM script: Each cell represents the number of identical (ID), matching (MA) or similar (SI) items between the two related forms. Red cells show more than five comparable items. The complete table is available in the Supplement (Supplement file 3).

Fig. 4. Grid-Image illustrating the percentage of matching data elements from comparison of all urology forms with ADT, ONDIS and registries. Yellow cells represent high percentages of matching items; red cells represent low percentages of matching items between the two respective forms. The black boxes highlight high number of matching items between documentation forms: Box A demonstrates that the urology forms from Erlangen bear resemblance to ADT; Box B highlights the matching concepts between CDE_prostate and other forms; Box C shows the percentage of matching concepts between urology in Erlangen and ONDIS; Box D highlights matching items between the three included registry forms (EpiNRW, GEKID, Finish_registry) and the other forms within the comparison. – Further graphical results are provided in the Supplement (Supplement file 4, grid images figure III.1–7, dendrograms figure IV.1–4). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
in Fig. 7 illustrates which forms offer high number of comparable items with the different registries. Furthermore, the chart demonstrates that there are only few comparable items with Finish_registry within gynecological documentation. Comparing gynecology routine documentation with research documentation, Fig. 6 (part C) illustrates the comparison of routine documentation with NCT00490139. Further comparisons are shown in Fig. 7. The documentation set from Münster shares 24 and Erlangen 18 comparable items with NCT00490139. Further results of comparing breast cancer documentation with quality management, registries and research documentation can be found in the Supplement (Supplement file 5).

4.2. Analysis of documentation standards

The comparison of common data elements in prostate cancer (CDE_prostate) with the whole ADT basis documentation set delivers 12 matching and 8 similar items. The DPKK dataset shares 10 similar and 3 matching items with CDE_prostate. Comparison of CDE_prostate with registries and quality management forms is visualized in Fig. 5.

Analyzing the 30 most frequent comparable items within the comparisons of CDE_prostate with other forms dealing with prostate cancer, results in 26 shared medical and 4 administrative concepts. The full list is available in the Supplement (Supplement file 7). 38% of the most frequent medical concepts are within DPKK_prostate, 65% within ADT dataset, 23% within Onkozert_prostate and 46% within registry documents. The top 26 medical concepts are in 35% within forms from Münster and 100% of them are within Erlangen’s documentation.

Comparison of registries with each other and other documentation standards or datasets is visualized in Fig. 9. Comparing EpiNRW with Finish_registry results in 6 matching and 12 similar items. The 30 most frequent comparable items within comparison of registries with routine documentation, research documentation, certification and quality management forms for both entities are listed in the Supplement (Supplement file 8).

The result of comparing clinical trial documentation with other documentation standards using the example of the included trial NCT00490139 is illustrated in Fig. 10. The complete table with all forms that contain comparable items with NCT00490139 can be found in the Supplement (Supplement file 9).

Analysis of the top 30 most frequent comparable items between NCT00490139 and other forms for documentation of breast cancer results in 26 medical and 4 administrative concepts. The whole list is provided in the Supplement (Supplement file 10). 27% of the most frequent medical concepts are within AQUA_breast, 88% within ADT dataset, 23% within required items for Certification_breast and 62% within registry documents. 50% of the top 26 medical concepts are within forms from Münster and 46% of them are within Erlangen’s documentation set. Further results of analyzing documentation standards can be found in the Supplement (Supplement file 5).

4.3. Common data elements for prostate and breast cancer

4.3.1. Prostate cancer

Analyzing and comparing forms for clinical documentation of both hospitals with forms containing special items for prostate cancer (ADT_prostate, CDE_prostate, DPKK_prostate,
Fig. 6. Dendrogram illustrating the number of matching items within comparison of gynecology forms with special breast cancer forms: ADT breast items, AQUA breast surgery and certification for breast cancer. Forms with high number of matching items are clustered. The number of matching items is plotted on the y-axis. There is a high number of matching items between forms (A) two documents within one clinic, forms (B) certification and quality management forms, forms (C) clinical forms and trial documentation, between forms (D) forms of different working steps, forms (E) cancer conference protocols at different hospitals. Further graphical results are available in the Supplement (Supplement file 6, grid images, figure I.1–8, dendrograms, figure II.1–3).

Fig. 7. Spider chart illustrating the number of comparable data elements from comparison of registries (EpiNRW, Finish registry, GEKID), certification & quality management forms (Certification_breast, AQUA_breast, ADT_breast) and trial documentation (NCT00490139) with gynecology forms from Erlangen and Münster. Gyn-E_documentation-conference and Gyn-E_short-medical-history share at least 16 comparable items with EpiNRW. The Finish_registry shares barely comparable items with other forms. Larger numbers of comparable data elements are within the comparison with forms from Erlangen. Gyn-MS_conference-after-surgery and Gyn-MS_surgery-planning share up to 15 comparable items with the AQUA_breast. In comparison with Certification_breast especially Gyn-E_pathology-report and Gyn-MS_conference-after-surgery share comparable items.
Fig. 8. Spider chart illustrating the number of comparable data elements from comparison of CDE_prostate and DPKK_prostate, two forms from research documentation, with ADT dataset, the group of registries and Onkozert_prostate. CDE_prostate contains comparable items especially with DPKK_prostate and vice versa. ADT_diagnosis and EpiNRW share up to 13 comparable data elements. Comparison of one form with itself is not illustrated.

Fig. 9. Spider chart illustrating the number of comparable data elements from comparison of cancer registries (EpiNRW, Finish_registry and GEKID), with quality management forms (ADT dataset, AQUA_breast, Certification_breast, Onkozert_prostate), research documentation forms (DPKK_prostate, CDE_prostate) as well as trial documentation (NCT00490139). EpiNRW shares comparable items especially with GEKID and ADT_diagnosis. GEKID presents a similar pattern of comparisons. In contrast, Finish_registry provides overall a lower number of comparable data elements. Comparison of one form with itself is not illustrated.
Onkozert_prostate) allows to determine lists of common comparable data elements. 15 medical concepts were identified within the top 30 common concepts. The top 5 of those concepts are shown in Table 1. The whole list is shown in the Supplement (Supplement file 11).

53% of medical concepts are part of DPKK_prostate, 67% of CDE_prostate, 20% of ADT_prostate and 40% are part of certification requirements Onkozert_prostate. The common concepts are found in up to 16 of 39 forms (general cancer documentation and special prostate cancer forms).

Within the 15 most frequent common medical concepts there are several items from pathology, items with information about treatment and special items for prostate cancer. 56% of the determined concepts are part of documentation forms in Münster and 100% of the items are within Erlangen’s documentation set.

Table 1
Top 5 medical common concepts determined within documentation forms dealing with prostate cancer. The whole list is available in the Supplement (Supplement file 11). Within the top 5 are concepts regarding diagnosis (“date of diagnosis” and “Diagnosis ICD code”), one special concept for prostate cancer (“Prostate-Specific Antigen”) as well as a concept about therapy and extent of disease.

<table>
<thead>
<tr>
<th>Name</th>
<th>Concept code (UMLS CUI)</th>
<th>In number of forms</th>
<th>In% of forms</th>
<th>Routine documentation</th>
<th>Registries</th>
<th>Certification &amp; quality management</th>
<th>Research documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of diagnosis tumor finding</td>
<td>C1274082 C2316983</td>
<td>16</td>
<td>41%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Diagnosis ICD code of tumor finding</td>
<td>C1274082 C2598420</td>
<td>16</td>
<td>41%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td>C0392920</td>
<td>10</td>
<td>26%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Prostate-Specific Antigen</td>
<td>C0138741</td>
<td>7</td>
<td>18%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Site of distant metastasis</td>
<td>C0807944</td>
<td>7</td>
<td>18%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 2
Top 5 common concepts determined within documentation forms dealing with breast cancer. The whole list of top 30 items is available in the Supplement (Supplement file 12). Within the list of the top 5 concepts there is one special concept for breast cancer (“Her2/Neu Status”), a further pathology concept (“Number of examined lymph nodes”) as well as concepts regarding therapy planning.

<table>
<thead>
<tr>
<th>Name</th>
<th>Concept code (UMLS CUI)</th>
<th>In number of forms</th>
<th>In% of forms</th>
<th>Routine documentation</th>
<th>Registries</th>
<th>Certification &amp; quality management</th>
<th>Research documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planned operative surgery</td>
<td>C0543467 C1301732</td>
<td>11</td>
<td>31%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Number of examined lymph nodes</td>
<td>C2733494</td>
<td>11</td>
<td>31%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Planned pain management</td>
<td>C0002766 C1301732</td>
<td>10</td>
<td>28%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Clinical Trial</td>
<td>C0008976</td>
<td>9</td>
<td>25%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Her2/Neu Status</td>
<td>C1512413</td>
<td>8</td>
<td>22%</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
4.3.2. Breast cancer

Lists of common comparable data elements were determined by form analysis regarding clinical documentation in gynecology of both hospitals with special forms for breast cancer (ADT_breast, AQUA_breast, Certification_breast) as well as the included clinical trial documentation (NCT00490139). Within the top 30 common concepts 14 medical concepts were identified. The top 5 concepts are shown in Table 2. The whole list of common concepts is shown in the Supplement (Supplement file 12).

57% of medical concepts are part of AQUA_breast, 43% of ADT_breast, 50% are part of certification requirements Certification_breast and 36% part of trial documentation for NCT00490139. The common medical concepts are found in up to 11 of 36 forms (general cancer documentation and special breast cancer forms).

Within the 14 most frequent medical concepts there are items from pathology as well as clinical parameters (e.g. “Menopause status” or “Palpatory diagnosis”) and information regarding treatment. The items from pathology are mainly in the breast cancer documentation. 100% of the determined concepts are within Münster’s documentation set and 93% within the forms from Erlangen.

5. Discussion

Documentation is really an extensive task and characterizes the work in today’s medicine. The amount of documentation for patients with a cancer diagnosis is enormous. For a patient with prostate cancer within one hospital there are up to 1125 data elements to document, including routine documentation, cancer registry and ADT as quality management. If the patient is included in a trial there is an additional documentation burden. This finding is consistent for two very common cancer diseases and two German university hospitals. So there is an urgent need for standardized structured single source documentation. The high number of data elements raises also the question whether all these data elements are important and needed. This is underlined by the fact that both hospitals differ in their documentation forms. To answer the question concerning relevant elements, unified documentation sets with common concepts have to be developed and agreed upon.

Up to our knowledge it was the first time that such analyses based on a great number of real forms and requirements were performed. The same medical concepts were found in up to 16 forms, so there is a huge redundancy of documentation. This indicates opportunities for standardization. With the proposed approach it was shown that structured and semantic enriched documentation forms can be compared automatically to identify common data elements in medical documentation. Our analysis demonstrates significant similarities between documentation for routine, quality management and research in two major cancer entities at two sites. This finding indicates a high re-use potential and the need for structured documentation with coded items.

The heterogeneity of forms in our analysis demonstrates the necessity and importance of common data element collections. Therefore, it is important to strive for improved standardized documentation sets based on common data elements.

Our analysis applies and extends the compareODM method [18]. This method requires two coded forms in CDISC ODM format as input and identifies identical, matching, similar and differing data elements.

To enable analyses of large form sets it was necessary to enhance compareODM (see Section 3.3). We applied this new approach to analyze the documentation landscapes of two very common diseases considering routine documentation at two sites, different cancer registries, certification & quality management documents and research documentation. A lot of overlaps and variations between different documentation modalities were identified. Precise lists of common concepts were provided to improve design of health information systems.

To our knowledge, similar analysis approaches for medical documentation are not available in the literature.

5.1. Limitations

Our analysis was conducted in the German healthcare setting. We also took into account international common data elements, an international trial and a registry from Finland. However, it would be interesting to replicate this study in other countries.

Regarding the collection of forms it has to be mentioned, that from Erlangen’s gynecology department only four forms were included, which were available within the HIS. These forms do not cover the whole process from admission to discharge.

Forms were coded with regard to the medical concepts and a similar structure for common data items was chosen. Nevertheless, there are a lot of differing items between forms. One reason is the structural difference of forms, such as conversion of paper forms into electronic ones. Some paper forms are ambiguous, for instance with unclear data types or combined elements (for example checkbox + string).

In addition, there is a varying level of detail regarding documentation. Registry documents are mostly more general, for example they contain a general item for TNM “T stage”.

The clinical documentation in contrast contains either a “clinical T stage” or “pathological T stage” item. The comparison results in zero comparable items because the current version of compareODM does not consider relationships of semantic codes.

5.2. The role of semantic annotation

Using semantic annotation is feasible and necessary for systematic analysis of forms. In addition, semantic annotation is a necessary to detect the different types of similarity: identical, matching and similar items. Especially similar items could enable automated secondary use of data and avoid redundancy if they were standardized.

Coding of forms by adding medical concepts is a very complex, time-consuming and difficult task. As Zhu stated “one of the primary, yet most fundamental, challenges in exchanging and integrating data is to ensure that data is both semantically (i.e., variable names and values share common meanings) and syntactically (i.e., the data shares a common format) interoperable” [14]. Unfortunately, to our knowledge there is no official guideline for coding data elements.

We chose UMLS as code system for our analysis to reach the highest coverage of coded items because it contains a lot more concepts than other systems (e.g. UMLS contains about 3 Mio concepts [37] SNOMED about 300,000 concepts [38]). The group of SNOMED Codes is a subgroup of UMLS. To demonstrate the differences, we performed a form comparison for two forms (Gyn-MS_checklist vs. Gyn-MS_medical-history) based on UMLS and SNOMED coding (see Table 3). The number of items without code is obviously

<table>
<thead>
<tr>
<th>Type</th>
<th>SNOMED</th>
<th>UMLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Matching</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Similar</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Differing</td>
<td>31</td>
<td>49</td>
</tr>
<tr>
<td>Without code</td>
<td>25</td>
<td>1</td>
</tr>
</tbody>
</table>
higher if only SNOMED CT codes are used due to the lower number of available codes. The number of different items is lower due to the lower number of coded items. All in all, we were able to code 3417 items with UMLS and 2856 items with SNOMED codes from 3565 items in total. This means for UMLS a coverage of 96% and for SNOMED of 80%.

We encountered several coding issues: The description of clinical items is often general so that medical knowledge is needed for coding. There is often more than one concept within the code system and further effort is needed to find the appropriate code. Problems of choosing the right codes become obvious by comparing the chosen codes of two persons. Therefore, an intercoder comparison was performed with a medical expert and a linguistic expert. The result was even worse than expected as there was only a very small overlap within this coding. The analysis showed differences regarding postcoordination and indicated that comparison with compareODM depends on data structure and data type of items. Neglecting these categories, the comparison provides about 23% of the codes within the forms as equal. Further analysis showed that the coding approach of both experts was different. The medical expert searched for codes that represent the medical concept in the best way and the linguistic expert focused on item names and searched for codes to transform it into codes, for instance:

- “Count of involved lymph nodes” (Count C0750480, Invasive C1517574 and lymph nodes C0024204 vs. Number of lymph nodes involved by malignant neoplasm C2732750).

According to the result coding by physicians is preferable. The discovered effect was already reported by Andrews et al. [39] who stated that “individuals with different interests or training may use the same terminology to code the same concepts differently”. Within his analysis only 14% of the coded data “were classified as computationally equivalent”.

We compared our work with results from Andrews et al. [40] and Chiang et al. [41]. They examined that three coders found the same core concept in up to 33% of the analyzed cases [40] respectively up to 53% (depending on the used browser for coding) “exact coding” with “complete agreement” [41]. Our rate of agreement is lower because the automated comparison with compareODM is more restrictive than a manual comparison. Overall, the need for standardization efforts was emphasized by Zhu mentioning that “heterogeneity in data representation, need for standardization efforts was emphasized by Zhu”.

5.3. Comparison results

5.3.1. Comparison of hospitals

A literature search was performed with the following keyword combinations: “documentation form comparison hospital”, “clinical documentation comparison”, “documentation cancer registry comparison”, “documentation cancer quality management comparison” and “clinical documentation form comparison” which could not identify similar work.

According to our analysis, substantial differences regarding documentation were identified between two university hospitals. While one department performs the whole documentation process within the HIS, another department works in large parts with paper forms. Considering reuse and reduction of workload, paper-based documentation has to be eliminated [45].

5.3.2. Prostate cancer documentation

The structured digital documentation in the urology department of Erlangen demonstrates that this approach works in daily life. Most matching items in the comparison of urology in Münster and Erlangen are within the cancer conference forms. This indicates that standardization of processes leads to harmonization of forms. The comparison of biopsy forms demonstrates that a pathology report can be a structured document and this should be applied in other clinical parts, where the pathology report is still a free text.

The whole documentation set in Erlangen is based on the ADT [46], therefore there is a high amount of matching items between items in Erlangen and ADT.

Comparisons with the ADT result in 161 comparable items within Erlangen’s documentation forms and 39 in Münster’s. On the other hand the two sites share only 56 comparable items. Building forms up on the basis of standardized datasets is successful and favorable regarding reuse. There are comparable items of several forms with the cancer registry EpiNRW. This shows that there might be a chance for automated reports, because urology forms from Erlangen contain almost 50% of the used concepts.

5.3.3. Breast cancer documentation

In Münster’s gynecology department there are a lot of comparable items within forms for different work steps (e.g. Cyn-MS_conference-after-biopsy, Cyn-MS_checklist-after-biopsy). This shows that it would be possible to have pre-filled items. Furthermore, there is great intersection of Cyn-MS_checklist and Cyn-MS_medical-history – forms that are needed for different purposes. This indicates an opportunity to save time by reusing information that is already documented in another form.

The cancer conference forms of both gynecology departments, as well as Cyn-MS_checklist-after-biopsy (Münster) and Cyn-E_registro-conference (Erlangen) have several comparable items. The second part reveals that there may be different work steps, whereas the same items are documented and needed.

Comparing routine documentation of the two departments with ADT dataset there are up to 53 comparable items per clinic. On the other hand the two departments share 58 comparable items. This indicates that building up documentation forms on the basis of standardized datasets is successful, especially if one considers the result within Erlangen’s urology.

5.3.4. Summary: Documentation for both entities

The small overlap (about 25% of all items) of the departments regarding their routine documentation shows the necessity and the potential for improvements. This fits with aim number eight from the German national cancer plan that demands a consistent dataset for cancer documentation [47]. It also demonstrates that standardization of datasets is a national task.

5.3.5. Cancer registries

The comparison of epidemiological registry in North-Rhine-Westphalia (EpiNRW) and the Finish registry (Finish_registry) with the CDE dataset (CDE_prostate) shows that the cancer registry covers international common elements that are based on scientific analysis. EpiNRW shares 13 and Finish_registry shares 6 comparable data elements with CDE_prostate. Furthermore, the overlap of items between the German and Finish registry (18 comparable...
data elements) shows that cancer registries are internationally established. The high number of comparable items between Epi-NRW, GEKID and ONDIS dataset indicates a close link between epidemiological and clinical cancer registries.

5.3.6. Research documentation

The comparison of CDE_prostate and DPKK_prostate shows, that there are comparable items in research forms from different countries. Comparing CDE_prostate with ADT, there is only a little number of comparable items in relation to this big dataset. This might be explained by different healthcare systems or different design goals.

Comparing single forms with NCT00490139 provides up to 40 comparable items. Structured documentation of adverse events by using Common Terminology Criteria for Adverse Events (CTCAE) – Criteria [48] is responsible for this high number of comparable items. Forms for documentation of chemotherapy and radiotherapy present the highest overlap with trial documentation.

Half of the determined top 26 most frequent used medical concepts are contained in Münster’s documentation set and 46% in Erlangen’s documentation. It must be taken into account, that the determined top items represent about 10% of the included items and the included items represent only 40% of the available trial forms. All in all it is shown, that structured documentation of adverse events enables data reuse for trials. El Fadly et al. found 13.4% of data elements from electronic case report forms (eCRF) within EHR templates as possible candidates for pre-population [12].

5.3.7. Reuse potentials

During the analysis we discovered some achievements which could be quickly realized. There are a few forms respectively data elements which are already available in the HIS and could easily be integrated in the current documentation, e.g. structured items within Münster’s cancer conference protocols (TNM stage). Adding these additional items will increase the matching items (7) and similar items (4) in comparison with Epi-NRW.

In addition, there are text modules actually used within Münster’s gynecology cancer conference forms that could be transformed to structured items and improve the possibility for reuse. The usage of the text modules content as structured items leads to a greater number of comparable items.

For example, regarding Gyn-MS_cancer-conference there are 13 additional comparable items in comparison with ADT dataset. The whole list of comparisons is shown in the Supplement (Supplement file 13).

In Erlangen these are the well-structured forms that are not used at the moment: Uro-E_chemotherapy, Uro-E_final-report, Uro-E_pathology-biopsy. The advantage of these forms from Erlangen is shown within the results. The mentioned forms are highlighted as forms with many comparable items. Especially Uro-E_chemotherapy is predestinated for reuse with 52 comparable items in comparison with ADT_chemotherapy. All in all, with minor changes greater impact on reuse can be achieved. The chances of structured documentation regarding re-use are evident.

5.4. Common data elements

Analysis was focused on forms designed for cancer diseases in the respective departments and therefore only the specific ADT forms (e.g. ADT_prostate) were included in the analysis of common data elements.

Analyzing the list of determined common data elements within forms for prostate cancer shows that 50% of the top 30 concepts are administrative items. Considering medical concepts there are several items from pathology (e.g. “pT category”, “Perineural invasion”, etc.), time designations (like “date of diagnosis”) and items for treatment information (e.g. “chemotherapy” or “repeated surgical procedures”). When analyzing common concepts within gynecology forms, especially 14 medical concepts are important items about pathology, clinical and treatment information. These identified common concepts are important and used in up to 16 forms.

Every item of our urology common data element collection is within the documentation set from Erlangen. The gynecology in Erlangen covers 13 of 14 medical common concepts and the gynecology in Münster 14 of 14. Analyzing the overlap of CDE_prostate and the determined common data elements for prostate cancer 10 of 15 (67%) common concepts are part of CDE_prostate. This indicates that the most important information about patients with prostate cancer is covered by the determined top 15 medical concepts.

Our approach of determining common data elements is different to the approach of Weintraub et al. [49]. They used a team of experts to determine key data elements of cardiovascular vocabulary. In our analysis we used statistics to determine common data elements. We examined the number of forms in which a medical concept was used and identified the top 30 concepts.

5.5. Future work

With this work we provide a proof-of-concept that automated comparison of large forms and identification of common concepts is feasible. Further analysis has to be done with focus on the specific data elements with data types and value lists, because we focused on the concepts.

In this context similar and matching items can be examined in particular to work out concrete suggestions for harmonization of the considered forms, thereby contributing to less duplicate data entry. The presented approach is promising and should be extended to other sets of forms that should be harmonized. In future there should be a holistic approach for development of documentation forms involving all relevant stakeholders, so that there will be more harmonized datasets and less redundant documentation. Avoiding duplicate data entry is not solved with the actual algorithm, but the presented concept of form comparison can contribute to solve this important problem.

6. Conclusion

Identifying common data elements in medical forms from different settings with automated and systematic comparison of forms is feasible.

Analyses within gynecology documentation provide up to 38% comparable items with the epidemiological cancer registry of North-Rhine Westphalia. Urology documentation shares up to 34% comparable items with ADT quality management forms. About 67% of the determined top 30 common data elements in urology are part of the international common data elements by CPCTR.

Authors contributions

AS and JT provided forms and supported clinical interpretation. HD supported analyses within the clinical cancer registry. JH provided ideas for the project. TB provided forms from Erlangen. RK coded forms. RK performed the analyses. BB and RK made adaptations of compareODM and added new features. RK wrote the manuscript. BB supported writing. BB, MD critically reviewed the manuscript.
Acknowledgments

We appreciate the permission to use the forms of the urology and gynecology department in Erlangen and Münster. We appreciate the permission for using CPTCR CDE for this project. We appreciate the permission for using the dataset of ONDIS for analysis.

Henning Siebel supported coding and the intercoder comparison. Katja Hebestreit supported programming in R. John Dik supported English writing.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2014.04.008.

References


