Availability of Structured Data Elements in Electronic Health Records for Supporting Patient Recruitment in Clinical Trials

Albert Vass\textsuperscript{a,b}, Ines Reinecke\textsuperscript{c}, Martin Boeker\textsuperscript{d}, Hans-Ulrich Prokosch\textsuperscript{e}, Christian Gulden\textsuperscript{e}

\textsuperscript{a} Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany
\textsuperscript{b} Data Integration Center, University Medicine Greifswald, Greifswald, Germany
\textsuperscript{c} Carl Gustav Carus Faculty of Medicine, Center for Medical Informatics, Institute for Medical Informatics and Biometry, Technische Universität Dresden, Dresden, Germany
\textsuperscript{d} Institute of Medical Informatics, Statistics and Epidemiology, University Hospital rechts der Isar, Technical University of Munich, Munich, Germany
\textsuperscript{e} Chair of Medical Informatics, Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

Abstract
Automated identification of eligible patients for clinical trials is an evident secondary use for electronic health records (EHR) data accumulated during routine care. This task requires relevant data elements to be both available in the EHR and in a structured form. This work analyzes these data quality dimensions of EHR data elements corresponding to a selection of frequent eligibility criteria over a total of 436 patient records at 10 university hospitals within the MIRACUM consortium. Data elements from demographics, diagnosis and laboratory findings are typically structured with a completeness of 73\% to 88\% while medication as well as procedures are rather unstructured with a completeness of only 44\%. The results can be used to derive suggestions for data quality improvement measures with respect to patient recruitment as well as to serve as a baseline to quantify future developments.

Keywords:
Data Accuracy; Electronic Health Records; Patient Selection

Introduction
The process of patient recruitment in clinical trials is crucial for the advancement of medical knowledge, yet time-intensive if done manually. It often leads to delays or the termination of clinical trials due to an insufficient number of participants [1]. However, electronic screening tools based on the secondary usage of electronic health records (EHR) data are a promising approach to speed up the process of patient recruitment [2]. Literature shows that data quality of the EHR data needs to be evaluated in order to utilize it for a purpose other than the original [3]. As stated by Weiskopf et al. one of the most important dimensions of data quality is its completeness [4]. For the use case of supporting patient recruitment, relevant data elements are those that are queried as part of a trial’s eligibility criteria. In this work, we investigate the availability and completeness of relevant data elements for patient recruitment within the EHRs of 10 different university hospitals in Germany that belong to the MIRACUM consortium of the Medical-Informatics Initiative.

Methods
Preparation
In our previous work, we identified parameters relevant to the patient recruitment use case based on their frequency of occurrence in eligibility criteria in a sample of 50 clinical trials [5]. From these parameters, the 28 most frequent were selected. This frequency is further referred to as relevance.

Data acquisition
Based on these selected parameters, a data extraction sheet was created to assess their quality in the EHR. The data extraction sheet was designed as a Microsoft Excel spreadsheet. It included meta-information of the case and patient documentation and is composed of data elements with case-specific information (e.g., admission date) as well as data elements on parameters from the data group of demographics, diagnoses, procedures, medication, laboratory findings, and scores. For each data element meta-information (e.g., documentation quality, date of first occurrence, date of first/last documentation) were requested for entry. For procedures and medication, the stages of history and stay, additionally, admission and discharge were recorded separately.

At each of the 10 participating sites, the 5 most relevant departments were selected based on the number of visits. From the selected departments, 9 EHR records from closed cases of in-patient stays between 3-15 days during the period from 1/1/2018 to 12/31/2018 and having a primary diagnosis in ICD-10 (GM) chapters C, E, G, or J were randomly sampled. Randomization was normalized to the frequency of diagnosis chapter occurrence, meaning that differences in the frequency of the different ICD chapters were balanced.

A total of 436 cases meta-information (approx. 45 per site) was transferred to a REDCap project [6, 7] for data consolidation. Depending on the site case documentation meta-information was entered directly to REDCap, or utilizing R [8] for the data transfer from the extraction sheet to REDCap. Minor inconsistencies that arose (invalid entries or similar) were resolved and corrected manually with the individuals involved in this step.
The evaluation was performed with R on an export from the REDCap project.

Analysis

Completeness

Focusing on the suitability of EHR data elements for patient recruitment, the first step was to determine how frequently a data element was present. For this, the decision if a data element has to be taken into account in a specific case was evaluated from meta-information (concordance): if at least one date meta-information is entered for a data element, it is taken into account, otherwise not. Data elements in which the date meta-information is only entered for the first occurrence are treated separately. If this is longer than 14 days before the date of admission, the corresponding data element is ignored, since it is most likely an element from a case in context of an earlier inpatient stay.

Proportion of structured data elements

A single data element is scored as structured (or unstructured), depending on whether its documentation quality meta-information has the value "structured" as well as "both" (or "unstructured"). Missing values for the data structure information (e.g. NA) get handled as unstructured data elements. The proportion of structured data elements for a particular parameter (or data group) is calculated with the mean function over all data elements that belong to the parameter (or data group). A breakdown by site and department is possible, but is not the focus of this analysis.

Results

The relevance as eligibility criterion versus relative completeness in the EHR systems is shown in Figure 1. For clarity, the data items are grouped into the data groups demographics, diagnoses, procedures, medication, scores, and laboratory and sorted vertically based on mean relevance (more relevant groups at the top). The data structure is divided into the two classes “structured” and “unstructured” and visualized by a color-code (less than 50 % structured: light; 50 % or more structured: dark).

Demographic data

This data group consists of the data elements date of birth and gender and has a relevance of 1.00. With a few exceptions, the group is completely available in a structured form. One site did not provide any information on the data elements, so they are scored as not available there. For another site, the completeness is 0.76/0.78 (date of birth/gender), resulting in an average completeness of 0.87 for this group.

Diagnoses

This data group (mean relevance 0.76) includes admission diagnosis, primary diagnosis, secondary diagnosis, and non-coded secondary diagnosis. The first three items are present at all sites; only 2 of 10 sites did not provide information on the non-coded secondary diagnosis. On average, this results in 0.81 for the degree of completeness, with the coded diagnosis ranging from 0.91 (secondary diagnosis) to 0.97 (primary diagnosis).

Procedures

This data group includes elements for coded and non-coded procedures, in each of the stages of history and stay. The distribution is rather inconsistent and broad, both within the different stages and data elements and between sites. Procedure history (coded) is present in only 3/10 sites. In contrast, procedure history (not coded) is present at 8/10 sites. The opposite picture emerges during the stay: then procedures (not coded) are present at 7/10 sites, while 10/10 sites have coded procedures. The degree of completeness is summarized in Table 1.
Table 1: Averaged proportion of present data items (completeness) from data group procedures

<table>
<thead>
<tr>
<th>case status</th>
<th>coded</th>
<th>not coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>history</td>
<td>0.15</td>
<td>0.48</td>
</tr>
<tr>
<td>stay</td>
<td>0.85</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Medication

This data group is composed of information on coded and non-coded medication, the dosage and the end date of medication for the 4 stages of history, admission, stay, and discharge. Medication is documented very heterogeneously, there are significant differences both between sites and individual data elements.

Table 2 - Averaged proportion of present data items (completeness) for type of medication

<table>
<thead>
<tr>
<th>case status</th>
<th>coded</th>
<th>not coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>history</td>
<td>0.13</td>
<td>0.43</td>
</tr>
<tr>
<td>admission</td>
<td>0.20</td>
<td>0.51</td>
</tr>
<tr>
<td>stay</td>
<td>0.33</td>
<td>0.56</td>
</tr>
<tr>
<td>discharge</td>
<td>0.28</td>
<td>0.61</td>
</tr>
</tbody>
</table>

As an example, the proportion of existing items for (non-) coded medication is shown in Table 2. It is evident that there is a focus on non-coded medication documentation, although there are clearly outstanding sites (4/10) that document medication coded (e.g. ATC) with above-average frequency. Similar to procedures, the proportion of coded documentation versus non-coded documentation increases during in-patient stay.

Scores

This group includes the TNM, ECOG, and RECIST scores. The documentation quality is rather heterogeneous. Two of 10 sites do not supply information on any of those items. TNM and ECOG are typically present at most (8/10 and 7/10, respectively) sites (fraction of present items of 0.24 and 0.22, respectively), whereas RECIST (present at 2/10 sites) is the exception (fraction of present data items 0.10).

Laboratory

This group includes five typical laboratory parameters. These are recorded at all sites and the degree of completeness is on average between 0.61 and 0.87. The data are predominantly available in a structured form.

Discussion

In this study, the focus is on the suitability of the reused data for patient recruitment, which is why the structure of the data elements plays an important role. The demographics data group consists only of the two data elements date of birth and gender and is typically recorded almost completely and in a structured form. Still, in most cases, these parameters alone are too non-specific for the purpose of patient recruitment, and the remaining eligibility criteria have to be taken into account. Billing-relevant data elements are also recorded frequently and in structured form (diagnoses and procedures coded in each case).

Data elements that can be assigned to the laboratory findings data group have a high degree of structure, as they are usually fed directly into the EHR without manual interaction. The sites seems to use slightly different compositions of laboratory parameters in routine care so the completeness determined depends on how well the set of requested laboratory parameters is covered at each site.

For the data groups procedures and medication different stages (case status) have been recorded separately. Since the overall completeness shown in Figure 1 is averaged over the stages, it is in each of these two data groups distributed over a comparatively wide range. Nevertheless, there appears to be a systematic difference in completeness when comparing the history, admission, stay, and discharge stages. As shown in Table 1 and Table 2 for procedure code and medication code, respectively, the completeness during stay exceeds that at other stages.

Similarly, diagnoses include non-coded secondary diagnoses in addition to coded diagnoses, which contribute differently to quality due to non-coded recording.

Comparison with related work

Köpcke et al. [10] addressed a similar question in their work, although their approach slightly differed from ours. In a study selection, a mapping to patient characteristics was performed. These could be categorized into semantic groups according to Luo et al. [11]. The mapping to EHR data elements was done in a further step. For some patient characteristics, multiple EHR data elements were required for coverage. For completeness of topic groups, both availability of EHR data elements in the EHR and frequency of filling were considered.

A similar issue was also considered in the work of Doods et al. [9]. There, the EHR was used as the starting point for the structure of the data elements to be considered, and the selection was based on appearance in a sample of clinical trials. This makes the grouping of data elements very comparable to our grouping. Since the focus in the work of Doods et al. was feasibility, the question of whether a data element is structured or unstructured was not addressed separately.

Despite differences in approach (grouping of data elements) and orientation (focus feasibility vs. recruitment), our results can be compared with those of Köpcke et al. and Doods et al. A simplified overview of this is shown in Table 3.

A comparatively high completeness of demographic data is found in all three papers. The same applies to diagnosis, with Köpcke et al. reporting the frequency of a coded diagnosis aggregated with that of other data elements in the "health status" topic group. Similarly, procedures and medication are aggregated under "treatment and healthcare". The indication of completeness is compatible with the values of Doods et al. and those from this work. It is striking that in the present study differences in the completeness of the coded medication or procedures were found, depending on the stage of the patient's stay at which the documentation is made. Differences can be identified in laboratory findings. Köpcke et al. grouped other data elements under the topic group "diagnostic & lab test" that appeared to contribute lower completeness. The differences between Doods et al. (2011 data collection) and this work (2018 data collection) may stem from the fact that further development of the integration of laboratory systems with EHRs has led to an improvement.

A simplified overview of this is shown in Table 3.

A comparatively high completeness of demographic data is found in all three papers. The same applies to diagnosis, with Köpcke et al. reporting the frequency of a coded diagnosis aggregated with that of other data elements in the "health status" topic group. Similarly, procedures and medication are aggregated under "treatment and healthcare". The indication of completeness is compatible with the values of Doods et al. and those from this work. It is striking that in the present study differences in the completeness of the coded medication or procedures were found, depending on the stage of the patient's stay at which the documentation is made. Differences can be identified in laboratory findings. Köpcke et al. grouped other data elements under the topic group "diagnostic & lab test" that appeared to contribute lower completeness. The differences between Doods et al. (2011 data collection) and this work (2018 data collection) may stem from the fact that further development of the integration of laboratory systems with EHRs has led to an improvement.
Limitations

During the evaluation, it became clear that the nature of the data collection left discretion in completing the data extraction sheets for the processing staff and that this resulted in inconsistent approaches to data collection. For example, no validation was provided for the entries in the data extraction sheet, so a meta-information field to flag case-relatedly irrelevant data elements was used inconsistently across sites. This effect was countered during the analysis so that bias could be minimized.

The correctness of the available data elements was not the subject of this study and cannot be directly tested with the available data.

Conclusions and future work

There are clear differences in documentation quality of the assessed parameters. While data on patient demographics and conditions is both relevant and available, improvements in the documentation of medication and procedures could directly benefit efforts to support patient recruitment.

It is intended to analyze the temporal availability of the data elements in future work.

Acknowledgements

The authors would like to acknowledge the support of the entire MIRACUM Use Case 1 team.

This research has been conducted within the MIRACUM consortium. MIRACUM is funded by the German Ministry for Education and Research (BMBF) under the Funding Number FKZ 01ZZ1801A-M.

The present work was performed in (partial) fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.” from the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (CG).

References


Address for correspondence

Albert Vass <albert.vass@uni-greifswald.de>