Frequency Analysis of Medical Concepts in Clinical Trials and their Coverage in MeSH and SNOMED-CT*

J. Varghese; M. Dugas
Institute of Medical Informatics, University of Muenster, Muenster, Germany

Keywords
Eligibility criteria, CUI, data items, clinical trials, ODM, UMLS, MeSH, SNOMED-CT

Summary
Background: Eligibility criteria (EC) of clinical trials play a key role in selecting appropriate study candidates and the validity of the outcome of a clinical trial. However, in most cases EC are provided in unstandardised ways such as free text, which raises significant challenges for machine-readability.

Objectives: To establish a list of most frequent medical concepts in clinical trials with semantic annotations. This concept list contributes to standardisation of EC and identifies relevant data items in electronic health records (EHRs) for clinical research. The coverage of the list in two major clinical vocabularies, MeSH and SNOMED-CT, will be assessed.

Methods: Four hundred and twenty-five clinical trials conducted between 2000 and 2011 at a German university hospital were analysed. 6671 EC were manually annotated by a medical coder using Concept Unique Identifiers (CUIs) provided by the Unified Medical Language System. Two physicians performed a semi-automatic CUI code revision. Concept frequency was analysed and clusters of concepts were manually identified. A binomial significance test was applied to quantify coverage differences of the most frequent concepts in MeSH and SNOMED-CT.

Results: Based on manual medical coding of 425 clinical trials, 7588 concepts were identified, of which 5236 were distinct. A top 100 list containing 101 most frequent medical concepts was established. The concepts of this list cover 25 % of all concept occurrences in all analysed clinical trials. This list reveals six missing entries in SNOMED-CT, 12 in MeSH. The median of EC frequency per trial has increased throughout the trial years (2000–2005: 8 EC/trial, 2011: 14 EC/trial).

Conclusions: Relatively few concepts cover one quarter of concept occurrences that represent EC in recent studies. Therefore, these concepts can serve as candidate data elements for integration into EHRs to optimise patient recruitment in clinical research.

1. Introduction
An appropriate set of eligibility criteria (EC) is crucial for the design of meaningful and efficient clinical trials. Many studies have difficulties in patient recruitment due to the misinterpretation or false selection of EC, which results in not getting the required number of eligible study candidates or picking wrong study candidates, which can result in misleading study results [1, 2]. In most cases, a study protocol defines EC in free text. From a form-based view, EC can be seen as a set of Boolean data items, which refer to conditions a study subject has to meet (inclusion criteria) or must not meet (exclusion criteria) to be eligible for participating in a clinical trial. To identify these data items within a medical context unambiguously, one can annotate each data item by a semantic code identifier, which maps a name of a data item, written in plain text, to its unique medical concept. Thus, assuming an appropriate medical concept is available for each data item, semantic ambiguity due to homonyms or synonyms can be avoided when formalizing EC of clinical trials. As an example, an item can be named as ‘Length’. From a clinician’s point of view, this can refer to medical concepts like ‘Body Height’, ‘Arm Length’ or ‘Leg length’. By annotating this item with a semantic code, the medical context of that item is clearly defined and machine-readable, which leads to major benefits for data management or data analysis in clinical trials [3].

Our main goal is to obtain a list of most frequent medical EC-concepts with semantic annotations and to show to which extent such a list of only few elements can cover all EC-concept occurrences in clinical trials. If a considerable amount of EC-concepts can be covered, the list could serve as a guidance list for future designs of clinical trials to facilitate precise and computable EC and/or to optimise patient recruitment. Furthermore, this list could identify an important set of concepts to extend data items in Electronic Health Records (EHRs) to improve their interface with clinical research. It is not intended to provide details on how these concepts can be implemented into an EHR with specifications on the data structure of value domains since this work is going to focus on the concept domain. Nevertheless, some
suggestions for a working integration into EHR are addressed in the discussion section.

A large number of clinical trials have to be analysed to establish such a list. To enable feedback from the scientific community, all EC from analysed clinical trials are available with semantic annotations as medical forms on a forms repository: https://medical-data-models.org, called MDM. To get a clinical overview of our list of most frequent EC-concepts, we will cluster them within a hierarchical tree of clinical categories.

MDM is an open-access repository for medical forms that has been implemented in our institute. It stores a variety of different forms that are used in routine documentation and clinical research [4]. The central format of MDM for storing forms follows the specification of Operational Data Model (ODM) [5] by the Clinical Data Interchange Standards Consortium (CDISC). It allows the author of a form to annotate each data item of a form with semantic aliases such as Concept Unique Identifiers (CUIs) provided by the metathesaurus Unified Medical Language System (UMLS) [6]. CUIs uniquely identify concepts within UMLS. For more detailed descriptions on concepts and CUIs by UMLS, see [7]. The role of MDM in this paper is to store all analysed clinical trials as medical forms with semantic annotations and provide these to the scientific community.

Our approach to determine a list of frequent medical concepts for EC is new, since we follow a semi-automatic approach. The manual part is done by carefully annotating EC with semantic codes on the basis of medical experts and a large set of clinical trials. The automatic part is done on the basis of 1) string similarities of EC texts for additional suggestions to code revision of medical concepts and 2) counting semantic codes for frequency analysis of medical concepts. We compare our obtained list with two other lists of common data elements in eligibility forms, which were created by different methodologies. The first one is a list generated by Weng [8] that represents a purely computational approach applying Natural Language Processing (NLP) to 137,889 clinical trials. It consists of 115 medical terms. The second list is established by Doods [9] by manually analysing 17 recent clinical trials and retrieving 75 medical concepts, which are both frequent and available as data items in EHRs of several hospitals in Europe. Since both comparison lists contain around 100 elements, we will restrict our list to the 100 most frequent concepts and analyse aspects of concept coverage saturation by using further or fewer elements. Finally, we will compare the coverage of the top 100 concepts within two widely established vocabularies: 1) Medical Subject Headings (MeSH) [10] and 2) Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [11, 12]. Thus, we can discover missing medical concepts in both vocabularies, which are relevant in the clinical research domain, since they have been used frequently as EC in clinical trials.

Therefore, the contribution of this article is to:

- Determine the most frequent medical concepts in EC of clinical trials by applying manual semantic code annotation by medical experts, semi-automatic code revision and frequency analysis on the basis of a large set of clinical trials (n = 425).
- Provide these EC with semantic annotations to the scientific community via the MDM portal.
- Establish a manual hierarchical categorisation of those concepts by medical experts.
- Compare the coverage of these concepts in the vocabularies of SNOMED-CT and MeSH.
- Reveal missing relevant entries in the vocabularies.

2. Methods

2.1 Collecting Trial Forms

It was intended to analyse EC-concepts of clinical trials conducted at our local university hospital. All clinical trials, which have been conducted at the University Hospital of Muenster [13] from 2000 to 2011 were retrieved (n = 425) from ClinicalTrials.gov [14]. For every obtained trial ClinicalTrials.gov provides free text under the section ‘Eligibility’ on the web. EC of all trials were extracted manually. For each clinical trial the therapeutic area was determined based on expert decision (IV).

2.2 Semantic Annotation by Medical Coder and Coding Principles

In 2011, each clinical trial with its free-text eligibility criteria was formatted in CDISC ODM [5] and uploaded to the ODM-form repository MDM [4]. Thus, EC are expressed as data items in an ODM-form.

A data item contains the original EC-text, the medical concept of the EC and the appropriate CUI annotation. E.g. an EC string like ‘Age ≥ 18 years’ is expressed by a data item with EC-text ‘Age ≥ 18 years’, the EC-concept name: ‘Age’ and the CUI ‘C0001779’. Concept identification and CUI annotation was applied by one medical linguist with extensive coding experience acknowledged in [15]. For the frequency analysis of medical concepts we do not differentiate between inclusion criteria and exclusion criteria. To extract medical concepts and CUI annotation from free EC-text of clinical trials, the following coding principles were applied (see also Figure 1 that depicts the workflow of coding principles).

2.2.1 Discarding Non-selective EC

Discard an EC, which is not selective in patient recruitment, e.g. the EC ‘Gender: Both’ is not selective since it does not contribute to a meaningful filtering in patient recruitment.

2.2.2 Identification of Medical Concepts

Check if each single EC can be expressed by a single medical concept with sufficient clinical specificity based on manual expert’s review. If yes, retrieve the CUI of that medical concept by searching the UMLS metathesaurus. If an appropriate CUI cannot be retrieved via the metathesaurus, apply post-coordination (see below). If a single EC cannot be expressed by a single concept, apply decomposing (see below).
E.g. a single EC written as “Patient must be between the ages of 18 and 65 years” is mapped to the medical concept ‘Age’ (CUI C0001779) without coding the age interval since the given age interval is not a concrete medical concept but a part of a value domain. Temporal relationships were not taken into account. E.g. ‘Type 1 diabetes for at least 24 months was mapped to the medical concept ‘Diabetes mellitus, Insulin dependent’ (CUI C0011854).

2.2.3 Correctness and Maximum Specificity

If several concepts can be retrieved for one EC via the metathesaurus, the concept with highest clinical specificity according to expert’s review was selected. E.g. an EC-text: ‘Patient with type 2 diabetes’ is being mapped to a medical concept. Searching the metathesaurus retrieves concepts named: 1) ‘Diabetes Mellitus, Non-Insulin-Dependent’, 2) ‘Diabetes Mellitus’, 3) ‘Type 2 diabetes mellitus in obese’. 1) provides the only medical concept that represents the EC-concept correctly and with highest clinical specificity, 2) is too general, 3) too specific.

2.2.4 Decomposing EC into Different Medical Concepts

Many different medical concepts appearing within a single EC-text, which might be connected via semantic operators (e.g. ‘Not’, ‘With’, ‘And’, ‘Or’, ‘By’, ‘Due to’), have to be extracted and mapped to constituent concepts. The connectors were not mapped since only the frequency of medical concepts will be analysed. E.g. an EC stated as ‘pregnancy or breast feeding’ would be decomposed into the concepts ‘pregnancy’ (CUI C0549206) and ‘breast feeding’ (CUI C0006147). In case of EC text like ‘Diagnosis of disease X’ we made use of decomposing, since we deemed the process of diagnosing and the diagnosed disease itself as two concepts. Note, that this principle is only applicable if different medical concepts exist in the text of a single EC. It can’t be applied if different medical words contribute to one whole medical concept. E.g. ‘Malignant hyperpyrexia due to anesthesia’ (CUI C0024591) is a single medical concept and must not be decomposed into the concepts ‘Malignant hyperpyrexia’ and ‘Anesthesia’.

2.2.5 Post-coordination of a Medical Concept

For EC mapping, pre-coordinated CUI-Codes are used whenever possible. However, if a single medical concept can’t be mapped to a single CUI, post-coordination will be considered. That is, further CUIs were additionally added to define the medical concept. No specific post-coordination syntax (for instance SNOMED-CT or caDSR as described in [16]) is applied. Instead, all post-coordinated concepts codes are listed subsequently for one medical concept. Order of codes is arbitrary, e.g. the annotation: ‘C0678852 C0205411’ referring to the concept ‘adequate organ function’ is equivalent to the annotation: ‘C0205411 C0678852’. It has to be noted that applying these principles of post-coor-


2.3 Semi-automatic CUI Revision by Physicians

All 425 forms in ODM format were retrieved from the MDM database. Every EC-concept with its name (e.g. 'Age'), EC-text (e.g. 'Age ≥ 18 years'), CUI annotation and form number was inserted into an R-data frame for CUI revision and frequency analysis.

Semantic annotations of EC-concept were reviewed by two physicians by focusing only on whose names were word-similar. A string A is word-similar to string B if the whole ordered sequence of words in A is contained in B (ignoring case and special characters such as brackets, ‘!’, ‘.’, etc.). A word is any sequence of characters and multiple words are separated by a blank space. Analogously, B is word-similar to A if the whole ordered sequence of words in B is contained in A (order of words is not contained in the opposite pair element).

Examples for non-word-similar pairs are:

1) 'Pregnant' and 'Pregnancy',
2) 'Multiple brain damages' and 'Multiple brain metastases' (full ordered sequence of words is not contained in the other pair element),
3) 'Brain damage, severe' and 'Severe brain damage' (order of words is not the same).

Thus, for manual physician’s review, it is possible to focus on typical CUI coding mistakes caused by homonym-like expressions (word-similar words with different meanings, but the same CUIs were used) or synonym-like expressions (word-similar words with the same meaning, but different CUIs were used). The automatic process of listing word-similar text labels was achieved by taking the name of the first EC-concept into a list and adding all names of the remaining EC-concepts to the list, which are word-similar by using R-function grep [17]. This function supports sub word matches on the basis of regular expressions. This step will be applied again on the second element of the list and so forth until all concept names have been taken into consideration. Thus, a result list is generated containing EC-concepts with word-similar concept names, which will be presented for manual review. For the manual review, both physicians went through the result list and decided whether or not a code correction will be applied based on a mutual consent. Further information on word-similarity, alternative string similarity measurements and the resulting list of word-similar EC-concept names is provided in the supplement.

Manual corrections are divided into EC-Discarding, simple corrections and corrections by decomposing. EC-Discarding deletes concepts of an EC if that EC is deemed non-selective according to the coding principles. A simple correction is applied when the CUI of one concept is replaced by a different more appropriate CUI. A correction by decomposing is applied if further medical concepts had to be added to represent the EC text.

2.4 Establish a Top 100 List

Finally, by counting the occurrences of some CUIs the frequency of medical concepts is calculated. In case of post-coordination the concept comparison is applied, since there is no ordering in post-coordinated concepts codes according to coding principles. A graph illustrating the cumulative EC-concept coverage by using more or less medical concepts is generated.

2.5 Manual Categorisation of Medical Concepts by a Heat Map Tree

Based on the consensus of two physicians (JV, MD), all concepts of the top 100 list are assigned to hierarchical categories within a clinical context. Every concept of the list is manually assigned to its immediate parent node, which represents the immediate clinical category the concept is assigned to (e.g. "COPD" is assigned to its immediate parent node "Pulmonary diseases", which is linked to the parent node "Internal diseases"). Parent nodes are assigned to their parent nodes until a tree with a root node, that subsumes all categories, is established. For reasons of clarity the depth and density of the tree will be limited such that a node will only be added to the graph if it represents more than two medical concepts of the top 100 list. Heat map colouring of the nodes correlates to the number of concepts per category and therefore indicates its relevance in clinical trials.

2.6 Coverage in MeSH and SNOMED-CT

For every medical concept in the top 100 list, its coverage in MeSH (Version 2013_1_21) and SNOMED-CT (version 2013_31) is assessed using the UMLS metathesaurus (version 2013 AA). If one of the concepts of the top 100 list doesn’t have an entry in one of the vocabularies, the UMLS database will be searched for an alternative medical concept. This alternative concept is considered equivalent if the difference in clinical specificity is tolerable based on a manual review of physicians (JV, MD). E.g. the medical concept named as ‘Patient currently pregnant’ is annotated with its CUI C0549206, which is not linked to MeSH. However, MeSH defines a descriptor ‘Pregnant Woman’ with CUI C0033011, which is considered equivalent based on manual review. A counterexample is the concept named ‘Allogeneic Hematopoietic Stem Cell Transplantation’ with CUI C2242529. Again, no MeSH entry is linked to this CUI, the closest medical concept with an existing MeSH entry is ‘Hematopoietic Stem Cell Transplantation’, which refers to a more general medical concept but cannot be considered equivalent, e.g. allogeneic and ‘autologous stem cell transplantation are different procedures. Missing medical concepts will be listed explicitly in the results section in order to reveal relevant concept gaps in the mentioned vocabularies. A binomial sign test (paired test, exact version of the McNemar’s test [18]) is applied to test for significant differences in concept coverage between MeSH and SNOMED-CT at a significance level alpha = 0.05.
2.7 Comparison to Other Lists of Common Data Elements in Clinical Trials

The overlap with two different published lists is investigated. A list by Weng [8] contains 115 of the most frequent word tags within EC text from ClinicalTrials.gov based on a fully automatic NLP approach. Since this list only contains word tags and not medical concepts, a match is considered if one of the word tags can be mapped to one of our medical concepts based on manual review (JV). A further comparison of our top 100 list with a list by Doods [9] is conducted. This list was established manually by pharmaceutical experts and health informatics professionals. It represents a set of 75 medical concepts, which have been identified as frequent in clinical trials and available in European EHR systems.

3. Results

3.1 Thematic Orientation and Semantic Annotation by Medical Coder

Overall, 20 different therapeutic areas could be identified with 'Hematology', 'Neurology' and 'Cardiovascular Diseases' being the three most frequent. Complete frequency distribution of all therapeutic areas is available in supplement, EC-Counts.xlsx. The initial coding process of all 425 clinical trials resulted in 6,671 EC referring to 7,340 medical concepts, 5,466 of which were distinct.

3.2 Semi-automatic CUI Revision by Physicians

During physicians’ review 722 of all 7,340 concepts (10%) had to be corrected. 581 concepts were (8%) were edited by simple corrections and 123 (2%) by decomposing and 18 had to be discarded because their EC were considered non-selective. For details on how each concept was corrected, see CUIcorrections.xls, available in the Supplement. After this revision step, all clinical forms contain 7,588 medical concepts, 5,236 of which are distinct (5,466 distinct concepts before revision). Hence, a reduced number of distinct medical concepts were obtained, though some EC had to be decomposed. This reduction is a result of reusing CUIs for concepts that were considered equivalent.

3.3 Top 100 of Most Frequently Used Medical Concepts

Frequency of medical concepts was summarised for all 425 trials. One hundred and one medical concepts were identified as most frequent (identical frequency for concept 100 and concept 101, full list available in the supplement). Table 1 provides a table with the top 15 extract of the most frequent concepts showing for each concept its CUI and average form frequency (AFF), which states the average occurrence of one concept within one EC-form. The most frequent concept is "Age" (CUI C0001779) with a frequency of 465, which results in an AFF of 1.09 (overall frequency: 465, forms in total: 425). The AFF is greater than 1.0 since in some clinical trials "Age" was used both as an inclusion and exclusion criterion. The least frequent concepts share a frequency of 4. Hence, the rest of all distinct EC-concepts not appearing in the top 100 list share a frequency of 3, 2 or 1. Summing up all frequencies of concepts from the top 100 list, an overall frequency of 1,876 was obtained. Since the corrected set of all (not necessarily distinct) concepts consists of 7,588 elements our top 100 list covers 25% of all concept occurrences. See table allconcepts-pub.xlsx in the Supplement for the complete top 100 list.

Table 1

<table>
<thead>
<tr>
<th>Rank</th>
<th>Concept name</th>
<th>UMLS_Code (CUI)</th>
<th>AFF</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>C0001779</td>
<td>1.09</td>
<td>465</td>
</tr>
<tr>
<td>2</td>
<td>Informed consent</td>
<td>C0021430</td>
<td>0.44</td>
<td>187</td>
</tr>
<tr>
<td>3</td>
<td>Patient currently being pregnant</td>
<td>C0549206</td>
<td>0.30</td>
<td>127</td>
</tr>
<tr>
<td>4</td>
<td>Breast feeding</td>
<td>C0006147</td>
<td>0.18</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>Diagnosis</td>
<td>C0011900</td>
<td>0.13</td>
<td>57</td>
</tr>
<tr>
<td>6</td>
<td>ECOG performance status</td>
<td>C1520224</td>
<td>0.12</td>
<td>52</td>
</tr>
<tr>
<td>7</td>
<td>Participation status in other recent studies</td>
<td>C2348568</td>
<td>0.09</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>Malignant neoplasms</td>
<td>C0006826</td>
<td>0.09</td>
<td>38</td>
</tr>
<tr>
<td>9</td>
<td>Life expectancy</td>
<td>C0023671</td>
<td>0.08</td>
<td>34</td>
</tr>
<tr>
<td>10</td>
<td>Drug abuse</td>
<td>C0013146</td>
<td>0.07</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>Pharmaceutical preparations</td>
<td>C0013227</td>
<td>0.06</td>
<td>27</td>
</tr>
<tr>
<td>12</td>
<td>Diabetes Mellitus, Non-Insulin-Dependent</td>
<td>C0011860</td>
<td>0.06</td>
<td>26</td>
</tr>
<tr>
<td>13</td>
<td>Alcohol abuse</td>
<td>C0085762</td>
<td>0.06</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>Contraception</td>
<td>C0700589</td>
<td>0.05</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>Creatinine measurement serum</td>
<td>C0201976</td>
<td>0.05</td>
<td>23</td>
</tr>
</tbody>
</table>

Figure 2a shows the cumulative coverage of all 45236 distinct concepts starting with the most frequent concepts "Age" (rank 1), "Informed consent" (rank 2), "Patient currently being pregnant" (rank 3) etc. on the left hand side of the graph. Figure 2b shows the cumulative coverage of the 200 most frequent concepts to focus on the left part of the graph. It can be seen that the coverage linearly increases by going beyond our list of 101 concepts. However, by using 101 concepts we already cover the steepest parts of that cumulative graph.

To summarise our key findings of this section: Based on expert-based semantic annotation of EC from 425 clinical trials, we could extract 5236 distinct medical concepts. By using 2% (= 101 : 5,236) of these distinct medical concepts, we can cover...
25% of EC-concepts occurrences in all analysed clinical trials.

3.4 Summary of Frequencies of EC and EC-concepts over Time

As mentioned above, EC were manually extracted from clinical trials. Since, some EC had to be decomposed into several EC-concepts, \(|\text{EC-concepts}| > |\text{EC}|\). In the following, the term \(\text{EC-concepts}\) is abbreviated as \(\text{concepts}\). After CUI revision, 6,671 EC were mapped to 7588 concepts. 2813 EC are inclusion criteria, 3858 are exclusion criteria. In terms of medians, a clinical trial consisted of 12 EC (Mean: 15.69, SD: 11.19) and 14 concepts (Mean 17.49, SD: 13.96). Figure 3 shows the median of frequencies of EC and concepts per trial from 2000 till 2011. It can be observed, that the frequencies have largely increased from 2000 to 2007. From 2007, medians of EC numbers barely changed, only a slight increase of concept numbers can be identified. The years 2000–2005 had to be pooled together, since only few trials were available from the corresponding constituent years on ClinicalTrials.gov. A detailed view on medians, mean numbers and standard deviations for every year, containing every study with its National Clinical Trial (NCT) number by Clinicaltrials.gov is given in ECCounts.xls in the supplement.

3.5 Semantic Tree of Medical Concepts

Figure 4 shows the semantic tree depicting categories and subcategories fitting the 101 most frequent medical concepts. The colour of a node indicates the number of medical concepts that fall into the category represented by that node, similar to a heat map. For a detailed view on which concept was assigned to which (most specific) category, see table allconcepts-pub.xlsx (column: Adjacent Parent Node) in the supplement section.

3.6 Comparison with Other Lists of Common Data Elements

Table 2 illustrates the overlap with a list of 115 common word tags by Weng [8], a fully automatic NLP approach, and with a list of 75 medical concepts by Doods [9], a manual approach based on analysis by pharmaceutical experts and health informatics professionals. The matches in Table 2 refer to the number of element matches within a pairwise list comparison. Two elements of different lists are considered a match if both elements refer to the same medical concept. This matching step is applied manually by a physician (JV). E.g. for the comparison of Doods and our list 18 elements of both lists reference the same medical concept. Regarding the list by Doods it must be taken into account that the list only contains EC concepts, which are considered as frequently used in clinical trials and available in European EHR systems. The highest number of matches could be identified between our list and Weng’s list. Among all of the three pairwise overlaps, our list provides the two biggest (36 and 18 matches).
detailed view of all pairwise comparisons is provided in the ▶Supplement in ListComparison.xlsx.

3.7 Coverage in MeSH and SNOMED-CT

Table 3 is a 2 × 2 contingency table, which counts how many of our most frequent medical concepts are available in SNOMED-CT and/or MeSH.

A higher coverage of medical concepts is provided by SNOMED-CT, but the difference to MeSH is not significant (p = 0.1094, two-sided binomial sign test). A list of all concepts, which are not covered by SNOMED-CT or MeSH is given in the supplement (allconcepts-pub.xlsx, sheet “Coverage”).

4. Discussion

4.1 Using Only 101 Medical Concepts and Integration into Electronic Health Record

Based on the cumulative coverage depicted in Figure 2a, we do not achieve complete saturation of coverage by only using 101 concepts, but still the steepest slope-changes of the graph are covered. Furthermore, by using a list of around 100 elements, we provide a similar amount of elements as in the comparison lists by Weng [8] and Doods [9]. It has to be noted that our list only contains elements that refer to the concept domain. A medical concept can only serve as a – not necessarily complete – description for an explicit data structure when incorporating our elements into EHRs. Therefore, successful integration into EHR for automatic patient recruitment needs further specifications on the data elements such as data type and some contextual information as well, e.g. measurement unit and timestamp (allowing for temporal relations of EC). Regarding the data type, the value domain should cover values like ‘null’ or ‘currently no information available’. This is important, since many of the Top 100 elements require too-specific patient’s information (e.g. ‘had stem cell transplant’), which is not to be documented at the current state of the patient. Moreover, a logical representation of EC (e.g. in propositional logic) is necessary for standardisation and evaluation of logical operators (e.g. ‘And’, ‘Or’, ‘Not’) and temporal information of EC (e.g. ‘existing Diabetes Mellitus Type 2 and arterial hypertension and insulin therapy for over six months’).

4.2 Comparison to Other Published Lists of Common Data Elements in Clinical Trials

As mentioned in the introduction, the list of Weng is the result of a fully automatic NLP approach while the list of Doods is a result of manual expert review. For the identification of frequent medical concepts, one potential problem with the approach of Weng is the semantic complexity of a single EC. Ross [19] manually reviewed the semantic complexity of EC in clinical trials by examining 1,000 random clinical trials from ClinicalTrials.gov as medical experts. A single EC was marked as semantically simple or complex. Among all comprehensible and selective EC almost 80% were semantically complex and could not be decomposed into simple EC. This poses a challenge to NLP approaches without using medical expert

Table 2 Comparison of other published list of common data elements. Doods: n = 75, Weng: n = 115, our list: n = 101, (n refers to the number of list elements)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doods – our list</td>
<td>18</td>
</tr>
<tr>
<td>Weng – our list</td>
<td>36</td>
</tr>
<tr>
<td>Doods – Weng</td>
<td>12</td>
</tr>
</tbody>
</table>
knowledge. As an example, the following EC text: “Men and women ≥ 50 years of age” (Trial id: NCT00637377) could be mistakenly linked to the concepts “Gender”, “Female Gender” or “Male Gender” by applying simple text processing and CUI mapping. As a human medical expert it is obvious that this EC text actually refers to “Age” as a concept. The specific gender is of no relevance, since both genders are included. Another example is the use of informal hints or examples within the EC text as following: “Suspected or diagnosed demyelinating disease of the central nervous system (e.g. multiple sclerosis or optic neuritis)” (Trial id: NCT00329303). Here, “multiple sclerosis” and “optic neuritis” could be mistakenly considered as EC concepts, though they are only disease examples for the actual concept. These are only few examples regarding challenges of NLP in processing EC text. In our paper we make use of our MDM portal that contains a reasonably high number of clinical trial forms (n = 425), where all of the retrieved clinical trial EC were manually annotated with CUIs.

In case of semantically complex EC (SCEC), we could rely on expert decision and, if necessary, apply the principle of post-coordination and decomposing, which may not reversibly infer the exact text of SCECs but uniquely annotates SCECs to make them amenable for frequency analysis of medical concepts. Thirty-six percent of our list elements can be mapped to list elements by Weng, marking a notable overlap. Furthermore, our list provides medical concepts with CUIs that define EC in an accurate and language-independent way. The process of patient recruitment for clinical trials can be improved by leveraging EHR data, but the availability of well-defined data elements in EHR is limited due to unstructured and non-coded clinical data in EHR [20]. Our list can identify new data elements in EHR systems in a machine readable way, since we make use of semantic annotations. Since machine readable concept codes support electronic patient recruitment [20, 21], clinical research can benefit from our work. Twenty-four percent of data elements published by Doods [9] match our list, which is still a notable overlap. It has to be considered that data elements in [9] have been identified based on frequency in clinical trials, data availability in European EHR systems and a smaller set of trials (n = 17). From our perspective, the CUI list in this manuscript should be representative since it was generated from a large amount of clinical trial forms and analysed through the computational methods mentioned above. Due to use of expert knowledge, we assume that our list can represent common

Table 3  Contingency table showing the coverage of the most frequent medical concepts in both vocabularies, n = 101

<table>
<thead>
<tr>
<th>Exist in SNOMED-CT</th>
<th>Do not exist in SNOMED-CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exist in MeSH</td>
<td>87</td>
</tr>
<tr>
<td>Do not exist in MeSH</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 4  Semantic tree categorising all 101 concepts. Every node is annotated with a count number in parenthesis. This count number refers to the exact number of medical concepts falling into that node’s category. Note that the count number of apparent’s node doesn’t need to be the sum of its child nodes’ count numbers, since the parent node can contain concepts, which cannot be assigned to any of its child nodes.
EC more accurately than NLP methods alone.

4.3 Course of EC- and Concept Frequencies over Time

Our findings on the increase of EC- and concepts frequencies per trial from 2001 to 2011 are in accordance of [22] and [23]. Results in [22] indicate an increase of procedures between 1999 and 2005, [23] states an increase of costs and documentation efforts after implementing the European Clinical Trial Directive in 2001 into UK law. From our perspective, an increasing EC frequency per trial can be problematic, since an overly complex EC set (median 14 EC-concepts per trial !) can make it difficult to identify appropriate study candidates and hamper – in case of positive study results – transfer of new therapeutic approaches into routine care.

4.4 Concept Coverage in MeSH and SNOMED-CT

Most of the concepts in the list are covered by MeSH and SNOMED-CT, with SNOMED-CT providing a non-significant higher coverage. But still, we revealed 6 relevant missing entries in SNOMED-CT and 12 in MeSH. The UMLS metathesaurus provided full coverage. MeSH only contains around 25,000 medical descriptors, SNOMED-CT more than 330,000 medical concepts. The non-significant coverage difference is anticipated, since our list contains concepts, which are frequently used for EC in clinical trials. Thus, most of concepts were expected to be available in both vocabularies.

4.5 Limitations

First, 425 clinical trial forms were semantically annotated with CUI codes by hand. Manually applied semantic annotation is a time-consuming process, since every data item has to be assigned to its medical concept as specifically as possible. Also, the manual part within semi-automatic CUI revision is time-consuming, which restricted the authors to only focus on word-similar EC-text to identify coding mistakes by the initial coder. Thus, this sort of CUI revision cannot identify a coding mistake of an EC whose text label is not word-similar to any of the other EC text labels.

Second, using the licensed UMLS metathesaurus for the retrieval of CUIs of medical concepts can lead to numerous valid search results, see our example above with ‘pregnant woman’ and ‘patient currently being pregnant’, both having different CUIs. This is due to the fact that UMLS is not a classification of concepts, it functions as a meta-vocabulary unifying big medical vocabularies but also trying to map same medical concepts into one CUI. However, this attempt of unifying same concepts from different vocabularies still needs further revision by UMLS. In our work, the issue of numerous valid search results for CUI code assignment could be significantly alleviated by applying expert judgement on the basis of defined coding principles and consensus of medical experts. Furthermore, we added a computational step to reduce coding mistakes due to word-similar concept names. E.g. an EC-concept like “informed consent” can be mapped to the CUI codes: 1) C0021430 (name: ‘informed consent’) and 2) C0184704 (name: ‘procurement of patient informed consent’). Both CUI codes are different but refer to the same medical concept. So if a medical coder had used the first code for annotating that concept in one trial and the second code for the same concept in another trial, our semi-automatic revision step would have identified both concept names as word-similar and presented both codes for expert review. Thus, reusing existing codes for same concepts can be facilitated. Discussion on other potential string similarity measurements is summarized in the supplement in CUIRevision.xlsx.

Third, all 425 clinical forms are related to one hospital: the University Hospital of Muenster, Germany (UKM). This potential selection bias could have an impact on the frequency analysis of medical concepts based on the study focus of that hospital. However, since all forms were collected from a long-time interval of more than 10 years and UKM is a hospital of tertiary care providing multidisciplinary clinical research and the given frequency distribution of therapeutic areas of our studies, we don’t believe that our trial selection process induces a significant bias producing misleading results.

4.3 Conclusion

We determined a list of most frequent medical concepts used in patient recruitment for clinical trials. The most frequent concept categories are related to diagnostic or disease information. The concept ‘Age’ belonging to the category ‘Demographics’ was used with the highest frequency. A set of 101 frequent concepts was sufficient to cover one quarter of all concept occurrences in EC of clinical trials. Therefore, these concepts can serve as candidate data elements for integration into electronic health records to optimise patient recruitment in clinical research.

References