Exploring Complexity in Health: An Interdisciplinary Systems Approach
A. Hoerbst et al. (Eds.)

doi:10.3233/978-1-61499-678-1-282

Key Data Elements in Myeloid Leukemia

Julian VARGHESE[a,1], Christian HOLZ[a], Phillip NEUHAUS[a], Massimo BERNARDI[b], Alexandra BOEHM[c], Arnold GANSER[d], Steven GORE[e], Mark HEANEY[f], Andreas HOCHHAUS[g], Wolf-Karsten HOFMANN[h], Utz KRUG[i], Carsten MÜLLER-TIDOW[j], Alexandra SMITH[k], Ansgar WELTERMANN[c], Theo de WITTE[l], Rüdiger HEHLMANN[m] and Martin DUGAS[a]

[a] *Institute of Medical Informatics, University of Muenster, Germany,* [b] *Hematology and BMT Unit, San Raffaele Scientific Institute, Italy,* [c] *Department of Hematology, Elisabethinen Hospital Linz, Austria,* [d]*Department of Hematology,Hannover Medical School, Germany,* [e]*Department of Internal Medicine, Yale Cancer Center, New Haven, CT, USA,* [f]*Department of Hematology, Columbia University Medical Center, USA,* [g]*Department of Internal Medicine, University Hospital Jena, Germany,* [h]*Department of Hematology and Oncology, University Medical Centre Mannheim, Germany,* [i]*Department of Hematology and Oncology, Klinikum Leverkusen, Germany,* [j]*Department of Hematology and Oncology, University Hospital Halle, Germany,* [k]*ECSG, Department of Health Sciences, University of York,UK* [l]*Radboud University Medical Centre, Nijmegen, Netherlands ,* [m]*Department of Hematology and Oncology, University of Heidelberg, Mannheim, Germany*

Abstract. Data standards consisting of key data elements for clinical routine and trial documentation harmonize documentation within and across different health care institutions making documentation more efficient and improving scientific data analysis. This work focusses on the field of myeloid leukemia (ML), where a semantic core of common data elements (CDEs) in routine and trial documentation is established by automatic UMLS-based form analysis of existing documentation models. These CDEs (n=227) were initially reviewed and commented by leukemia experts before they were systematically surveyed by an international voting process through seven hematologists of four countries. The total agreement score was 86%. 116 elements (51%) of these share an agreement score of 100%. This work generated CDEs with language-independent semantic codes and international clinical expert review to build a first approach towards an international data standard for ML. A first version of the CDE list is implemented in the data standard Operational Data Model and additional other data formats for reuse in different medical information systems.

Keywords. AML, CML, MDS, Common Data Elements, UMLS, Data Standards

1. Introduction

Medical documentation is time-consuming and non-uniform both in clinical routine [1] and clinical research [2]. Unstructured data capture and heterogeneity of different documentation models of health information systems increase redundant double documentation and hamper reuse of clinically relevant data. The documentation process

[1] Corresponding author: varghesejulian@gmail.com

is very complex, especially in oncology [3] due to a growing number of data elements, which are relevant for different medical purposes such as medical quality assurance, follow up information and clinical research.

Missing data standards result in heterogeneous but also incomplete or error-prone data capturing. Furthermore, data from patients coming from external health care providers cannot be transmitted quickly without a loss of data quality. Scientific data analysis is also complicated, because the primary data is captured in an unfavorable way (e.g. as free text in a long text field). Some approaches to propose key data elements within Electronic Health Records to unify clinical workflows and clinical trial documentation are recently being developed such as in acute coronary syndrome [4] or NIH-driven CDE initiatives [5]. Nowadays, establishing international data standards is a very laborious task involving tedious discussions and consent iterations among clinical experts [4]. Furthermore, existing registries or common data elements are implemented in different databases of different information systems with missing semantic annotations. The lack of semantic annotations leads to a lack of semantic interoperability, which can be a root cause of data sharing and integration problems among different information systems [6]. To our knowledge, data standards in the field of Leukemia with semantic enrichment by terminology codes do not exist.

Our objective is to collect existing documentation items for myeloid leukemia regarding clinical routine and trial documentation and to enrich them with semantic annotations. With this approach a semantic core of common data elements can be identified, standardized and presented to clinical experts. This method has a key advantage over a purely manual expert-driven consensus: Clinical experts do not necessarily have to participate in multiple tedious sessions starting from scratch. Instead, a well-filtered core set of key-data elements based on existing and/or successfully applied documentation models is generated, upon which experts can vote whether or not a data element should be integrated into an international standard within few sessions. Since our work also includes a systematic voting process by several international experts, our generated list of data elements can serve as a first draft for an international data standard.

2. Methods

2.1. Collecting relevant documentation models

Through our clinical partners within the European Leukemia Net (ELN) [7] we inquired Case Report Forms (CRFs) for studies on myeloid leukemia. Table 1 lists all clinical trials we received CRFs from by contacting the principle investigators. Since bone marrow transplant is a potential treatment for Leukemia and myelodysplastic syndrome (MDS) is known to progress to Myeloid Leukemia, registries of the European Society for Blood and Marrow Transplantation (EBMT) and the European MDS Registry (EUMDS) have been added to our documentation sources as well. Doehner et al., 2010 [11] published a set of recommendations for clinical routine management of Acute Myeloid Leukemia (AML), from which data elements could be extracted as well.

Table 1. Documentation sources [11] for the analysis of common data elements. CT = Clinical Trial, ST= Standard

	MDS	AML	CML
Doehner et al. 2010 ST		X	
HOVON 132 CT		X	
AML-AZA CT		X	
EUMDS registry	X		
CML-Tiger CT			X
EBMT registry	X	X	X

2.2. Common data element generation and clinical voting

All original forms, containing the existing data items of the six sources were recreated in the standardized format Operational Data Model (ODM), which is specified by the Clinical Data Interchange Standards Consortium (CDISC). All data items were defined by their names, datatypes and semantic codes according to the Unified Medical Language System (UMLS). Semantic coding was carried over by a physician (JV) according to published coding principles [8] using pre– and post-coordinated UMLS codes. The semantically enriched ODM forms were analyzed by CDEGenerator, an in-house implemented Java-based web-application, available online on [9]. CDEGenerator automatically sorts medical concepts (e.g. medication) of the existing data items according to their frequency (by counting identical UMLS codes) and also shows similarity of medical concepts based on code overlaps of post-coordinated concepts, e.g. “medication start date” is similar to “medication end date”, because the main concept “medication” is the same. An initial list of most frequent medical concepts is generated. By adding to each medical concept its datatype and possible values, e.g. code list items based on the majority of the referring existing data items, a medical concept also represents a data element. A data element will be added to our preliminary common data element set if it occurs at least twice within the six sources or if it is listed within the Doehner, et al. 2010 [11] standard. All data elements were initially reviewed and commented by eight leukemia experts and presented in a workshop in this year’s ELN conference to remove non-relevant/redundant or to add further medical concepts. The resulting list will be referred to as our CDE list, or as the key data element list. Apart from those eight experts, a web-based survey was implemented to invite further hematologists to systematically vote for every resulting data element whether or not a data element should be added to a potential international data standard for myeloid leukemia consisting of clinically relevant items for clinical routine and clinical trial research. Thus, every data element will be implemented into the survey as a question, including the data element name, its datatype and possible data element value sets. A vote for a specific data element is made by selecting one of four options: "Agree", "Unsure", "Disagree", and "Other: comment". The last option enables the participant to add further text, which will be collected for future review to update the CDE-List.

The agreement score for one data element is measured by the number of voters who selected “Agree” divided by the number of all voters. The agreement score for a medical category is the average of agreement scores of data elements in that category. The total agreement score (TotalAVG) refers to the average score of all data elements.

A last survey question enabled the participant to suggest further data elements not mentioned in the survey. Full instruction details of the survey are available on [10] and were presented to participants before starting the survey.

3. Results

After semantic code evaluation, 793 distinct data elements could be identified. Among those, 225 data elements met the frequency criteria described above. These were presented to an initial review of eight leukemia experts, after which three data elements were removed and five were added, thus the resulting CDE list contains 227 elements. All 227 CDEs were manually assigned to 14 medical categories, which are listed in Table 2. Seven further hematologists from four different countries (2x Austria, 2x Germany, 1x Italy and 2x USA) completed the survey. Each of the 14 medical categories formed the question groups of the survey. Table 2 summarizes for each manually identified medical category, its number of containing data elements and the average agreement score within that category. Figure 1 shows a graph depicting the medical categories, their agreement scores and their basic documentation order. Of course, the order can vary or some of the documentation nodes could be skipped among different patients. Lab measurements contribute to 61% of all CDEs representing the largest categories and also having the highest agreement scores except for Urinalysis, which has the lowest agreement score (18%) with low standard deviation among all voters (SD 7%). The total average agreement score for all CDEs was 86%. 163 elements (72%) of these have all an agreement score of more than 86% (at least six of seven raters voted "Agree"), 116 elements (51%) share an agreement score of 100% (all raters voted "Agree"). One participant proposed a further data element to be included for MDS: "WPSS Score". Full data containing all CDEs, their category and agreement votes of all hematologists are available on [11]. Additionally, a standardized data model of the CDE list has been implemented in ODM, available on [12] with additional other data formats (such as REDCap™ and HL7-CDA files) for reuse in different medical information systems.

Table 2. N: Number of data elements, CAS: Average category agreement score, SD: Standard deviation of agreement scores within one category, TotalAVG: Average Agreement Score for all data elements, Total SD: Standard deviation of agreement scores of all data elements, GVHD: Graft versus host disease.

Medical Category	N	CAS	SD
Administrative/Demographic details	11	0.90	0.13
Medical history	9	0.87	0.11
Physical examination/Follow up	14	0.76	0.2
Global disease course	11	0.79	0.18
Apparatus-based diagnostics	4	0.64	0.34
Lab: Blood panel	58	0.84	0.15
Lab: Urinalysis	7	0.18	0.07
Lab: Cytology/ Cytochemistry	14	0.82	0.24
Lab: Cytogenetics	14	0.94	0.11
Lab: Molecular Genetics	28	0.93	0.13
Lab: Immunophenotyping	17	0.99	0.03
Bone marrow transplant	13	0.95	0.09
GVHD	11	0.95	0.10
Treatment details	16	0.92	0.10
	Total: 227	TotalAVG:0.86	Total SD: 0.20

4. Discussion

Most of the generated CDEs share high agreement among hematologists. The category Medical history shows high agreement, but in clinical practice, medical information within this category tends to be documented heterogeneously and/or unstructured as free text [3]. Thus, data elements of our CDE list could immensely harmonize and structure

documentation processes across different health care institutions. The work has some limitations. AML is overly represented within the input sources compared to CML due to availability of CRFs. However, later clinical expert review including a question about missing data elements only indicated one missing element. Furthermore, the number of involved hematologists for the electronic voting process was limited (n=7) and we cannot draw any evidences on the awareness and attention every voter has payed to the questions. While the latter issue is a principal problem of all surveys, the first issue can be addressed by acquiring more hematologists to review. Since the web-survey is still active enabling clinicians to complete the survey anytime at any place, we will collect and analyze future survey data to improve validity and to keep the CDE list up to date in accordance to future requirements.

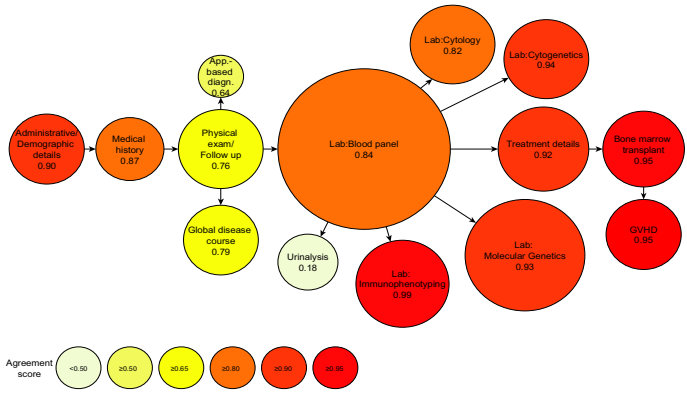

Figure 1. Basic documentation order of leukemia patients with the nodes representing medical categories of the common data elements. The area size of a node corresponds to the number of data elements in that category, the color to the average category agreement score.

References

[1] Ammenwerth E, Spötl HP. The time needed for clinical documentation versus direct patient care. A work-sampling analysis of physicians' activities. Methods Inf Med. 2009;48(1):84-91
[2] Getz K. Protocol Design Trends and their Effect on Clin. Trial Perform. RAJ Pharma May 2008, 315-316
[3] Krumm R et al. The need for harmonized structured documentation and chances of secondary use. J Biomed Inform. 2014 Oct;51:86-99
[4] Cannon CP et al. 2013 ACCF/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes and coronary artery disease. Circulation. 2013 Mar 5;127(9):1052-89
[5] NIH CDE Initiatives. Available from: https://cde.nlm.nih.gov/home. Accessed 2016 March 8
[6] Dugas M. Missing semantic annotation in databases. The root cause for data integration and migration problems in information systems. Methods Inf Med. 2014;53(6):516-7
[7] European LeukemiaNet. http://www.leukemia-net.org/. Accessed 2016 March 8
[8] Varghese J, Dugas M. Frequency Analysis of Medical Concepts in Clinical Trials and their Coverage in MeSH and SNOMED-CT. Methods Inf Med. 2015;54(1):83-92
[9] CDEGenerator Webtool. Available from: https://odmtoolbox.uni-muenster.de/CDEGenerator/CDEGenerator.html . Accessed 2016 March 8
[10] Websurvey CDE Leukemic Diseases. https://umfragen.uni-muenster.de/index.php/516388/lang-en. Accessed 2016 March 8.
[11] Sources CRFs and Doehner et al. (2010) standard. Survey Data. Available from: https://uni-muenster.sciebo.de/index.php/s/ytu4DnRQ7BOTJzU . Accessed 2016 March 8
[12] CDE Models. Available from: https://medical-data-models.org (Search Term: "Common Data Elements Myeloid Leukemia", Accessed 2016 March 12