


RESEARCH ARTICLE

Open Access



# Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting

Philipp Bruland<sup>1\*</sup> , Mark McGilchrist<sup>2</sup>, Eric Zapletal<sup>3</sup>, Dionisio Acosta<sup>4</sup>, Johann Proeve<sup>5</sup>, Scott Askin<sup>6</sup>, Thomas Ganslandt<sup>7</sup>, Justin Doods<sup>1</sup> and Martin Dugas<sup>1</sup>

## Abstract

**Background:** Data capture is one of the most expensive phases during the conduct of a clinical trial and the increasing use of electronic health records (EHR) offers significant savings to clinical research. To facilitate these secondary uses of routinely collected patient data, it is beneficial to know what data elements are captured in clinical trials. Therefore our aim here is to determine the most commonly used data elements in clinical trials and their availability in hospital EHR systems.

**Methods:** Case report forms for 23 clinical trials in differing disease areas were analyzed. Through an iterative and consensus-based process of medical informatics professionals from academia and trial experts from the European pharmaceutical industry, data elements were compiled for all disease areas and with special focus on the reporting of adverse events. Afterwards, data elements were identified and statistics acquired from hospital sites providing data to the EHR4CR project.

**Results:** The analysis identified 133 unique data elements. Fifty elements were congruent with a published data inventory for patient recruitment and 83 new elements were identified for clinical trial execution, including adverse event reporting. Demographic and laboratory elements lead the list of available elements in hospitals EHR systems. For the reporting of serious adverse events only very few elements could be identified in the patient records.

**Conclusions:** Common data elements in clinical trials have been identified and their availability in hospital systems elucidated. Several elements, often those related to reimbursement, are frequently available whereas more specialized elements are ranked at the bottom of the data inventory list. Hospitals that want to obtain the benefits of reusing data for research from their EHR are now able to prioritize their efforts based on this common data element list.

**Keywords:** Clinical trials, Common data elements, Data quality, Electronic health records, Metadata, Secondary use

## Background

Data collection is one of the most expensive processes during the conduct of clinical trials. Over the last decade the number of clinical trials and the size of trials have steadily increased [1]. Likewise, the number and complexity of case report forms (CRFs) capturing the data for trial subjects grew as well. From a hospital perspective, the use of

electronic health record (EHR) systems and consequently the number of patients having at least a basic electronic medical record has experienced a significant and steady growth [2]. The transition from paper-based to electronic documentation has resulted in clinicians spending around 25–30% of their time on electronic documentation tasks [3, 4].

Recent research has shown that a certain amount of EHR data elements are available and suitable for different research purposes [5–9]. Nevertheless, it is important to note that the provenance, availability, degree of

\* Correspondence: philipp.bruland@uni-muenster.de

<sup>1</sup>Institute of Medical Informatics, University of Münster, Münster 48149, Germany

Full list of author information is available at the end of the article



standardization and structure of the EHR data plays a major role in its re-use for clinical research purpose [10–12].

Currently, the exchange of routinely collected data between EHR systems and clinical research databases is not fully automated and requires human intervention. This manual step is time-consuming, error-prone and also demotivating [5]. Transferring data electronically from an EHR source into an electronic data capture (EDC) system in a systematic, auditable and unambiguous manner provides several advantages, avoiding the detrimental effects of repeated data entry, decreasing documentation time and improved data quality and cost-effectiveness [5, 7, 13, 14].

On-site data monitoring is an expensive process for pharmaceutical companies as well. Monitors have to visit all sites to perform the so-called 'Source Data Verification' (SDV) by comparing source materials at the sites with data that has been entered into the trial database, e.g. an Electronic Data Capture (EDC) system. This tedious, time-consuming and expensive process might be optimized through a connection between EHR systems at the sites and the EDC system. If such a link is established and presumed data are validated, SDV could likely be reduced or even be eliminated. Time-consuming site visits would be reduced to a minimum and the monitors could focus on other aspects of the clinical trial conduct.

The Electronic Health Records for Clinical Research (EHR4CR) project [15], which is funded by the Innovative Medicines Initiative (IMI) has investigated these potential incentives and benefits [16]. The project is a public-private-partnership consisting of 34 partners from European pharmaceutical industry and academic institutions. Clinical partners were from France, Germany, Poland, Switzerland, and the United Kingdom. The participating companies from the European Federation of Pharmaceutical Industries and Associations (EFPIA) were: AMGEN, AstraZeneca, Bayer Health Care, F. Hoffmann-La Roche Ltd, GlaxoSmithKline, Johnson & Johnson, Lilly, MERCK KGaA, Novartis Pharma AG and Sanofi-Aventis.

The project's aim was to develop methods and a software platform as well as an accompanying business model to support clinical trials based on routinely collected data from EHR systems. Addressed scenarios included 'protocol feasibility', 'patient identification and recruitment', 'clinical trial execution' (CTE) and 'serious adverse event reporting' (SAE). Disease areas which the project focused on were diabetes, cardiovascular, infectious, oncology, neurology and respiratory diseases. In addition to establishing the benefits associated with the first two scenarios, the 'business model' workgroup also showed substantial potential savings in the scenario of 'clinical trial execution' [17].

The EHR4CR net benefits are obtained by offsetting these savings against the expenditure for setting-up the infrastructure to allow the re-use of routinely collected clinical data. Suitable data elements need to be identified,

new structures for documentation procedures might have to be established, and dedicated exports from the source EHR to the research database have to be maintained. In order to control and reduce these operational costs, the best approach is to focus on the most frequently used data elements of clinical trials.

The ability to pre-specify common data elements (CDEs) would greatly improve the setup process of electronic databases and simplify the exchange of medical data between different systems, i.e. EHR and EDC systems. Subsequent analyses are then accelerated due to fewer data transformations from different sources, and enhance the comparability of outcomes. Additionally, CDEs might help to reduce the number of, and focus on, relevant data elements that should be captured across all and in certain therapeutic areas. These effects should be especially favorable in the context of multicenter trials that could benefit from a common data model. Several initiatives and research groups have tackled this issue and defined common data elements for different therapeutic areas [18–22]. Common data elements are defined as metadata information that is of interest or relevance in a specific research domain. As part of the EHR4CR project, common data elements for the scenarios of protocol feasibility and patient identification and recruitment have previously been determined by Doods et al. [23, 24].

In order to re-use CDEs for clinical trials, they must firstly be available for documentation in the EHR information systems and secondly must be actively used to contain patient data. Several research groups have examined the presence of clinical trial data elements within existing EHR systems and found a broad range of coverage between 13 to 70% [5–9]. However, these results were only investigated for one clinical trial [5–7, 9] and an analysis of the frequency of documentation has so far only been performed in very few cases (e.g. Botsis et al. [25]).

Nevertheless, it remains unclear what kind of data elements are most commonly required for the documentation of clinical trials and whether those elements are available and also captured across EHR systems.

### Objective

Work Package 7 (Pilots) of EHR4CR developed an inventory of relevant data elements as an important cornerstone for the development of a system facilitating the secondary use of routinely collected data for the subject documentation in clinical trial. An inventory also supports calculations for business modeling to estimate whether this approach is economically feasible. The aim of the present research is to determine what data elements are the most frequently used in clinical trial execution. An additional focus was on data elements supporting SAE reporting. The fundamental question is whether those

data elements are covered by European EHR systems and how frequently they have been captured.

## Methods

An iterative consensus-driven approach was chosen for creating the inventory of CDEs for CTE and SAE reporting as well as their completeness of documentation.

## Material

CRFs from clinical trials were collected from all participating EFPIA companies within the EHR4CR project to perform the analysis. Criteria for CRF selection were at least one comprehensive clinical trial including over 200 study locations and a four-digit planned patient enrollment number. According to the Good Clinical Practice guideline, CRFs are understood as printed, optical or electronic documents designed to record all of the protocol-required information to be reported to the sponsor on each trial subject [26]. A CRF consists of several forms, each one with a different purpose/domain (e.g. demographics, vital signs, adverse events, and also multiple instances of, sometimes unscheduled, follow-up forms). We analyzed 23 trials covering the following disease areas as listed in table 1: cardiovascular, diabetes, infectious, neurology, oncology, psychiatric and respiratory. The reviewed clinical trial forms ranged between 22 and 164 and contained in sum 1086 forms. We used all forms of a CRF for the analyses, also repeating and unscheduled ones. The most comprehensive clinical trial amounted to a total of 3581 data elements.

The CRFs were in different computer processable formats such as XML-based Operational Data Model from the Clinical Data Interchange Standards Consortium (CDISC) or Excel spreadsheet files. Apart from CDISC Operational Data Model, all files were confirmed as having different structures.

For the availability and completeness evaluation we analyzed data exports from seven hospital sites for the data inventory. Some sites provided data from hospital-wide EHR systems while others only data from subsystems or data warehouses, for instance specialized systems for

breast cancer or diabetes. For one site data was only available from in-patient cases.

## Methodology

After collecting the trial CRFs, medical informatics professionals from academia and trial experts from EFPIA in the EHR4CR project were involved in the consensus-driven process for creating the data inventory. Data exports were performed by different university hospitals across Europe to assess the availability of these distinct elements and their frequency of documentation.

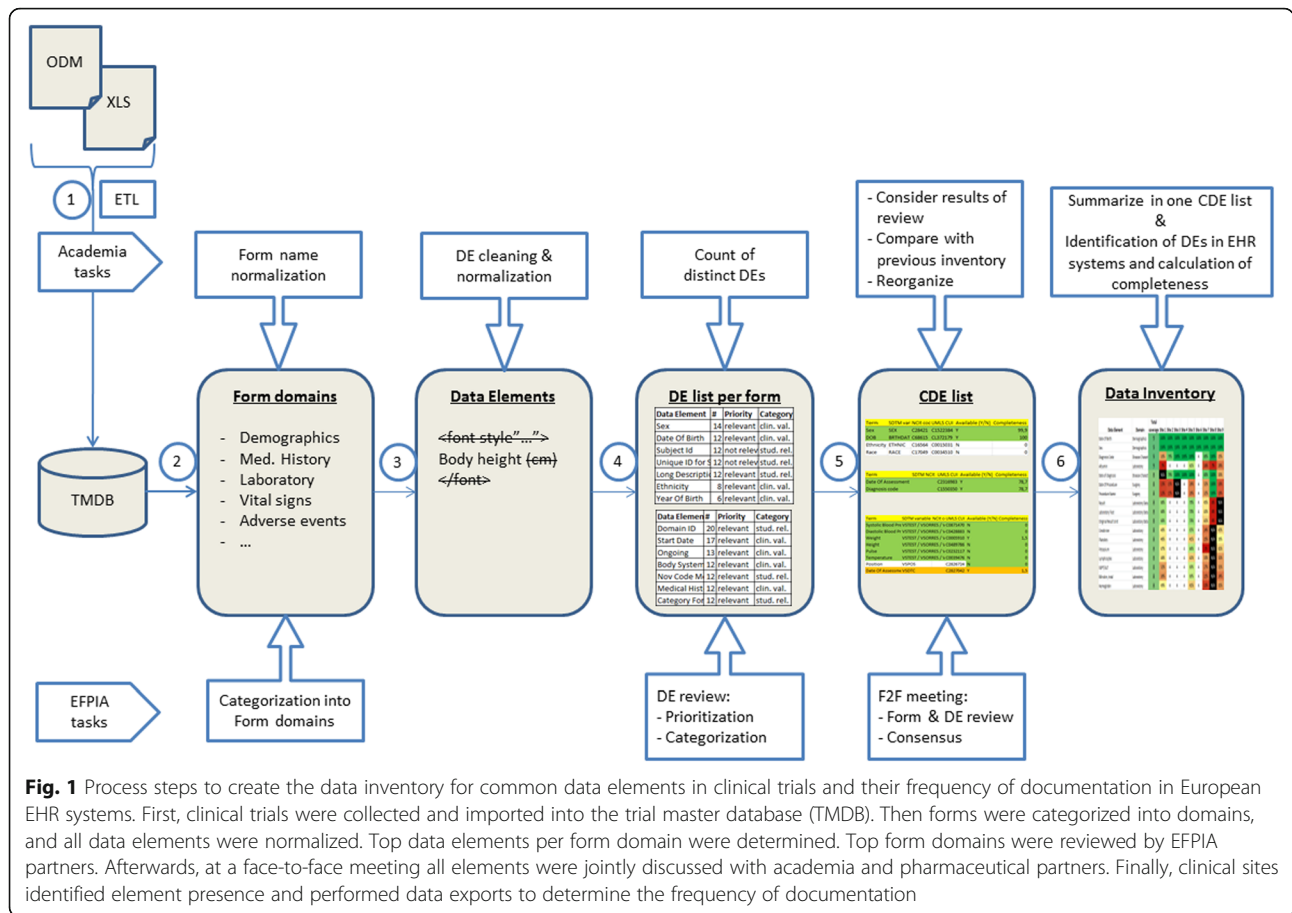
Figure 1 illustrates the process and involved parties who conducted the analyses for CTE and SAE reporting.

The following steps were performed:

1. Import: First, all collected CRFs were transformed and loaded into the TMDB (Trial Master Database) using the ETL-tool 'Talend Open Studio for Data Integration' [27]. The TMDB is used to gather all differently structured clinical trial CRFs together within one structured database. Meta-information for each trial is included concerning the disease area, planned patient enrollment numbers and participating sites. All CRFs and the associated data elements were imported into the TMDB.
2. Form categorization: To determine a top list of form, they need to be categorized by a domain, which was allocated by trial experts from EFPIA, each focusing on the trials provided by their company. Domains are understood as topic-specific classes (i.e. 'medical history', 'vital signs' or 'concomitant medications') that deliver additional contextual information to sites when exporting information on data elements they hold. Where possible, we assigned preferably the domains of CDISC's SDTM (Study Data Tabulation Model) [28], which is used for the definition and transmission of trial data. For instance, the form name 'Coagulation Panel' was allocated with domain 'Laboratory test results' or 'LB' in SDTM.
3. Data element normalization: During form categorization, data element names were normalized using phonetic and word similarity measures such as Levenshtein distance [29], Jaro-Winkler distance [30] and Metaphone [31]. This was a relevant process for the determination of common data elements. Special characters and HTML tags as well as style sheet information were removed using regular expressions. Measurement units were also removed since a conversion of values is feasible. If CDISC variable names were provided for an element, these names were additionally used for the normalization of element names. Finally, a manual review by trial experts was conducted to validate this part of the

**Table 1** Numbers of clinical trials and forms that were collected per disease area

Disease area	No. of trials	Forms
Cardiovascular	3	158
Diabetes	3	172
Infectious	2	60
Neuroscience	1	64
Oncology	3	192
Psychiatric	1	69
Respiratory	10	371
Sum	23	1086



process and determine results for those cases which could not be matched.

4. Data element review: Forms are used to collect data in clinical trials, therefore for the list of common data elements we chose to first rank the form domains by the frequency with which they appear in CRFs. We removed form domains that only appeared once. Then we calculated the frequency of each unique data element within each form domain. As a weighting factor we used the planned patient enrollment numbers we had available for each trial. There is an expectation that certain data elements will be found in certain form domains and not in others. Data elements within each form domain were independently reviewed by two EFPIA trial experts concerning an element’s relevance to the associated form domain. Duplicates were detected and CDISC variable names were added where possible. If data elements relevant to the domain were missing, the trial experts added them to the particular form domain. Distinctions between clinical parameters or administrative values (e.g. sequence numbers, subject or site identifiers) were also made to state whether data elements are

expected to be re-used from EHR systems. Due to the association between form domains and data elements, the frequency list for distinct CTE and SAE elements could be created.

5. Consensus meeting: As a last step, we refined the frequency lists in a face-to-face consensus meeting at which academic as well as pharmaceutical partners participated. The goal of this meeting was firstly to discuss and vote about the allocated form domains, data element names and semantic codes. It was stated whether a form domain should be kept in the overall list, whether an element should be removed in a form domain, and, if so, whether the naming was correct, whether it was a duplicate and finally which category (clinical parameters or administrative values) it belonged to. For instance, most elements that were stated as irrelevant were removed. After this consensus-driven step, the list was compared with the previous data inventory of ‘patient identification and recruitment’ to determine which elements had already been examined at the data provider sites. Some sites’ health information systems or data warehouses contained semantically annotated data elements. To support the discovery, mapping and data export

process, semantic codes were assigned to the elements in the list of the common data elements. Data elements of many CRFs already contained SDTM codes, so automated mappings to codes of the Unified Medical Language System (UMLS) were performed. Elements without SDTM codes were manually annotated by a medical informatics professional and a medical expert. SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms) codes were additionally added through a semi-automated identification process using the available mapping in the UMLS database.

6. Presentation of frequency list: Finally, the CDE list was distributed to the participating European university hospitals within the EHR4CR consortium. Data elements were located within structured documentation of their local data warehouses (previously populated with data from EHR systems), EHR systems or specialized subsystems and the frequency of occurrence was assessed, where possible. Sites with semantically annotated data sources programmatically identified matching entries between the CDE list and their source system based on identical results of code system codes. Frequency was calculated using the number of entered values which have been documented in the year 2013 divided by the total number of patients in 2013. To address privacy concerns and to obtain comparable values, relative percentages for a data element were

given. For instance, a frequency of 30.4% for 'Bilirubin, total' is the result of dividing the number of entered values (9574) by the number of patients in the year 2013 (31493). Where frequencies were not calculated in figs. 2 and 3, a distinction was still made between data elements that were 'available' (understood as possibility that the data element can be stored in the EHR system) and 'not available' for a given site. A heat map of data element frequency was created to depict the whole data element coverage, determined by the number of available elements. The heat map for SAE elements was reported separately. All analyses were performed using Microsoft Excel.

## Results

### Data inventory

After the consensus meeting 14 form domains remained relevant for potential pre-population in clinical trials. Table 2 shows the final form domains used in clinical trials with SDTM domain abbreviations where available and the frequency of occurrence, number of total forms and data elements that are contained in the respective domain.

Medical History, Adverse Events, Laboratory and Disposition are ranked at the top. Apart from Surgery, Physical Examination and Tumor Response all domains were present more than once in a clinical trial. SDTM domains

Data Element	Domain	No. of sites	Site 1 (DWH)	Site 2 (DWH)	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Occurrence in trials
Sex	Demographics	9	100%	100%	100%	100%	100%	100%	100%	100%	100%	25
Date Of Birth	Demographics	9	100%	100%	100%	100%	100%	100%	100%	100%	100%	34
Diagnosis Code	Disease Characteristics	9	33%	79%	100%	100%	100%	A	80%	100%	35%	12
Albumin	Laboratory	9	7%	A	A	A	61%	A	16%	7%	24%	10
Date of diagnosis	Disease Characteristics	8	N/A	79%	100%	100%	100%	A	80%	100%	35%	12
Result	Laboratory Data	8	68%	A	A	A	70%	A	45%	6%	N/A	8
Laboratory Test	Laboratory Data	8	68%	A	A	A	70%	A	45%	6%	N/A	8
Original Result Unit	Laboratory Data	8	68%	A	A	A	70%	A	41%	6%	N/A	8
Hemoglobin	Laboratory	8	49%	A	A	A	41%	A	13%	N/A	31%	14
Creatinine	Laboratory	8	48%	A	A	A	65%	A	24%	N/A	43%	17
Sodium	Laboratory	8	47%	A	A	A	65%	A	0%	N/A	41%	8
Potassium	Laboratory	8	47%	A	A	A	64%	A	3%	N/A	41%	10
Alkaline phosphatase	Laboratory	8	47%	A	A	A	61%	A	1%	N/A	24%	8
Protein, total	Laboratory	8	47%	A	A	A	5%	A	23%	N/A	18%	10
Lymphocytes	Laboratory	8	46%	A	A	A	41%	A	33%	N/A	31%	14
Platelets	Laboratory	8	46%	A	A	A	41%	A	25%	N/A	50%	14
Glucose, unspecified	Laboratory	8	40%	A	A	A	46%	A	3%	N/A	35%	17
Bilirubin, total	Laboratory	8	30%	A	A	A	60%	A	22%	N/A	24%	15

**Fig. 2** Extract from the entire data inventory for clinical trial execution and SAE reporting. On the left-hand side the data elements and their form domains are listed followed by the number of sites in which they occur and the sites availability. Site 1 and 2 used their data warehouse (DWH) for element identification and exports

Data Element	No. of sites	Site 1 (DWH)	Site 2 (DWH)	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Occurrence in trials
Date Of Death	7	1%	N/A	A	N/A	A	A	0%	A	1%	17
Start Date / Time	4	N/A	N/A	A	N/A	A	A	0%	N/A	N/A	34
Outcome	4	N/A	N/A	A	N/A	A	A	0%	N/A	N/A	34
Cause Of Death	4	N/A	N/A	A	N/A	A	A	0%	N/A	N/A	17
End Date / Time	3	N/A	N/A	N/A	N/A	A	A	0%	N/A	N/A	41
Verbatim Description	3	N/A	N/A	A	N/A	A	N/A	0%	N/A	N/A	38
Severity of Adverse Event	2	N/A	N/A	A	N/A	N/A	N/A	0%	N/A	N/A	7
Time Of Death	2	N/A	N/A	A	N/A	N/A	N/A	0%	N/A	N/A	5
Action(s) taken	1	N/A	N/A	N/A	N/A	N/A	N/A	0%	N/A	N/A	27
Seriousness of Adverse Event	1	N/A	N/A	N/A	N/A	N/A	N/A	0%	N/A	N/A	18
In Case Of Death, Autopsy Report	1	N/A	N/A	N/A	N/A	N/A	N/A	0%	N/A	N/A	7

**Fig. 3** Common data elements for the reporting of SAEs. More than half of the data exporting sites have no SAE documentation available in their EHRs. This refers to the items ‘seriousness’, ‘action taken’, and ‘autopsy report in case of death’. Some sites have few data elements available but unclear whether data values are present apart from one where no data has been captured so far

could be identified for all forms apart from ‘Surgery’ and ‘Disease Characteristics’. As shown in table 2 domains like ‘Demographics’, ‘Medical History’, ‘Adverse Events’ and ‘Vital Signs’ are used in all trials; ‘Tumor response’ only in one trial. The number of unique data elements which was the basis of these analyses ranged between 27 in ‘Tumor response’ and 2733 in ‘Adverse Events’.

After the consensus meeting the final data inventory was created which contains 133 data elements that are identified as the most frequently used elements in clinical trials. In a comparison with the previous data inventory for patient identification and recruitment 50 data elements are identical and the remaining 83 elements are new for the execution of clinical trials.

**Availability and frequency in European EHR systems**

Figure 2 presents an extract of the most frequently used data elements in clinical trials sorted by the frequency of captured elements in EHR systems. We differentiated between the presence (A=available; N/A=not available) and the frequency of captured data elements. Demographic and reimbursement data was ranked at the top of the list followed by several laboratory test results. The number in the right column indicates the frequency of occurrence in all forms of all trials.

The complete data inventory can be accessed in the additional material [see Additional file 1]. It also contains semantic codes of the UMLS and where possible of SNOMED CT as well as a definition of each data element.

**Table 2** Consensus list of top 14 form domains in order of frequency of occurrence. This list also includes the SDTM domain abbreviation if available and in how many trials the domain is used, the number of total forms and the containing data elements

Form domain	SDTM domain	In no. of trials	No. of forms	No. of data elements
Medical History	MH	23	85	1336
Adverse Event	AE	23	78	2733
Laboratory test results	LB	10	75	1643
Disposition	DS	19	74	653
Vital Signs	VS	23	56	974
Concomitant Medications	CM	22	52	1334
Questionnaire/Patient reported outcome	QS	16	35	1122
Demographics	DM	23	34	371
ECG	EG	12	30	436
Disease Characteristics	(ZC)	12	25	190
Substance Use	SU	14	22	246
Surgery	-	9	20	80
Physical Examination	PE	5	10	33
Tumor response	RS	1	7	27

In terms of reporting SAEs, fig. 3 shows the most frequently used data elements for the domain of SAEs.

Over half of the data providers reported not having any documentation structure to collect SAE information at all. Three reported that they have at least some elements available but it was unclear whether data has been collected in 2013. Site 7 has a complete electronic documentation for SAEs available within their EHR, but has never been used. Apart from the 'date of death', SAE related data has not been collected at all.

The complete inventory of data elements for 'clinical trial execution' and 'serious adverse event reporting' can be accessed online at: <https://www.medical-data-models.org/forms/17994>.

## Discussion

Re-use of routinely collected medical data is a promising approach to some of the problems clinical trials currently face. In most projects that attempt to use EHR data for purposes other than patient care, the re-use requires manual mapping between EHR and clinical trial data elements. To keep these efforts to a minimum, it is desirable to know which elements are most commonly used in clinical trials, to what extent such elements are available in EHRs and also to know how frequently such elements are currently documented.

The present research generated a list of common data elements found in clinical trials. It was compiled through an iterative and consensus-based process with medical informatics professionals from academia and trial experts from the European pharmaceutical industry. Through data exports performed at different university hospitals across Europe this list also presents the potential for re-use of EHR data in clinical trials. Different source systems, languages and terminologies underline the complexity of this research.

A large proportion of the data elements are laboratory analytes that are commonly required in clinical trials. These data elements are usually well structured within EHR systems so that these elements could be identified. During this localization process within the systems documentation structure it became apparent that some elements were present in multiple forms within the EHR. So, the origin and purpose had to be clarified also with the aid of the form domains and further semantic annotation. Especially the semantic annotation of data elements is a crucial but also one of the most labor-intensive and tedious tasks to be performed to facilitate the re-use of routinely collected healthcare data. A further issue was that several clinical concepts are only available in unstructured form as free text in the EHR. Most frequently available are demographic elements, captured by all hospitals followed by diagnostic and procedural entities. The remaining data elements are far less often captured, which is expected

since the main driver for data modeling in EHR systems is support of regulation, policy and reimbursement rather than clinical practice or research.

Since the CDE analysis did not focus on disease-specific data elements, exports by certain departments were not possible. Analyses for disease specific elements might have resulted in higher frequency values because disease-related elements are more frequently documented in their disease area. Some data elements are common between disease areas, whereas others do not belong to the particular subset. For instance, biopsy results would not exist in the cardiovascular disease area but rather in oncology, and conversely for laboratory results in cardiovascular disease. The site identification of these data elements and their associated data exports was a labor intensive and time consuming process for all participating sites. Conducting these analyses for elements in all disease areas would increase the workload substantially since data exports would need to be performed for each element in each area separately; dependent on how uniquely such elements are captured.

An additional focus of this work was to assess the coverage of data elements for SAE reporting. Not surprisingly, all clinical trials examined contained CRFs for adverse events. Our analysis showed that elements are highly standardized and also related to the SDTM domain of adverse events (AE). Nevertheless, apart from 'date of death' SAE elements such as start and end date, outcome, verbatim description, severity and seriousness or action(s) taken were not captured in the EHR system at all. This underlines that clinical practice and trial execution are different conceptual domains with different purposes for data capture.

However, it must be considered that some clinical phenomenon may not be directly collected as a data element but rather derived from different elements or different perspectives. For instance, 'cause of death' might be related to several finer grained data elements, such as biomarkers being completely out of normal range. Although they might be recorded, the 'cause of death' reported would be determined by a forensic pathologist, which could be very different from the perspective of a clinician. Hence, it might be worthwhile to investigate further the data structure of information systems, the purpose of data elements and the human rather than technical process of data collection.

For the development of the data inventory we chose to perform this process iteratively. This led us to a modified representation for the element list. In the previous EHR4CR data inventories only the frequency was stated. Instead, we decided to additionally indicate whether a data element is just available regardless of whether its frequency could be stated or not. A frequency of 0% would imply that the data element is present but data

were never collected. 'Available' gives the information that data might have been captured but could not be assessed.

### Related work

In the EHR4CR project, data inventories for 'protocol feasibility' [24] and 'patient identification and recruitment' [23] have been performed by Doods et al. There, 75 data elements were identified for feasibility assessment and 150 data elements for patient identification and recruitment. Despite the differing scenarios, a comparison with the current inventory for the execution and SAE reporting in clinical trials has shown that 50 data elements have already been identified and 83 are new data elements.

CDISC, C-Path, NCI-EVS and CFAST had introduced an initiative on 'Clinical Data Standards' to create industry-wide common standards for data capture in clinical trials to support the exchange of clinical research and metadata [32]. This initiative defines common data elements for different therapeutic areas. Currently, traumatic brain injury, breast cancer, COPD, diabetes, tuberculosis, etc. are covered. In addition, the CDISC SDTM implementation guideline contains a set of standardized and structured data elements for each form domain. The aim of this initiative is similar to ours concerning the identification of most frequently used data elements for clinical trials. Nevertheless, the focus of our work is different and goes beyond this initiative in terms of determining the availability and quality of data within EHR systems.

Köpcke et al. have analyzed eligibility criteria from 15 clinical trials and determined the presence and completeness within the partners EHR systems [33]. Botsis et al. examined the incompleteness rate of diagnoses in pathology reports resulting in 48.2% (1479 missing of 3068 patients) [25]. Both publications show that re-use of EHR data relies on the availability of (1) data fields and (2) captured patient values.

### Limitations

This research work aimed to build a data inventory for CTE and SAE reporting within the IMI funded EHR4CR project. Therefore, the inventory represents data elements that are important for trials conducted by European pharmaceutical companies as well as showing the availability and frequency within large European university hospitals. The number of clinical trials in this analysis is limited since most trial metadata are not publicly available [34] and EFPIA partners have provided a small number of clinical trials for the analyzed disease areas. So, this research could be treated as a pilot study and as a foundation for a more comprehensive analysis.

Data exports at the sites have been performed on different sources due to different site specific data access

policies: two sites queried their clinical data warehouses; four exported data directly from the EHR and because of permission restrictions two sites did not have the possibility to access patient data at all. One site was only able to export data from a dedicated system for one clinic. The time period for data analyses was the year 2013. At one site data for 2013 was not available in the data warehouse, so, they chose 2012 for their queries. Another site was only able to take the first half of the year for exports. It includes in-patients as well as out-patients (apart from one site) for all medical disciplines available in the source systems.

The data exports of this research represent only data of nine university hospitals across Europe. The generalizability of this approach relies on several aspects: First of all the adoption of EHR systems is a crucial indicator whether data could be made electronically available. The degree of digitalization is also a key factor. Although an electronic system is available it is not necessarily being used. Often, paper is still used in parallel. Last but not least, the degree of structuredness plays an essential role as to whether data is eligible for re-use.

Data elements in clinical trials are highly standardized and EHR documentation forms often contain unstructured information as free text, notwithstanding, initiatives or tools like openEHR [35] or Clinical Element Model [36] aim at defining standardized data elements for clinical documentation. Even though projects like SHARPN [37] or cloud4health [38] use natural language processing techniques for extracting relevant information from free text documents, the EHR4CR project did not focus on this approach.

### Further research

Several pharmaceutical companies have private data standards catalogs of forms which are used to create CRFs for clinical trials. Such catalogues ensure that a clinical trial doesn't have to start from scratch when setting up a new trial. In this regard, the pharmaceutical companies may benefit from each other when using a standardized catalog of data elements in certain areas together.

A further step for the EHR4CR project is the introduction of the CTE platform within the overall infrastructure. Consequently, the approach of electronic data exchange from patient care to clinical research has to be evaluated. It is also worthwhile to examine whether the application of the EHR4CR platform generates the desired advantages and savings for both sides: time saving and error-reductions for study nurses/sites and reduced SDV as well as more rapidly conducted data documentation for the industry or academic researchers. Reduced expenditure of time for SDV and data collection enables to focus more on other value added activities such as training and recruitment.



Similar to this work, most of the researches focus on the availability and quality of routinely collected patient data. However, it remains unclear whether the available common data elements are exactly the required elements for a clinical trial and whether the procedure of documentation satisfies the needs. For instance, the timeliness and the relationship between data elements play essential roles. The identification of suitable EHR data elements have been performed in all conscience, but the context of documentation remains at times unknown and is also controversially discussed in the literature [10, 39].

Further investigations concerning the data quality and the purpose of documentation are essential to ensure that correct data elements are selected for the secondary use of patient care data, in particular for a clinical trial execution, and that their contribution toward monitoring of therapeutic efficacy, patient-safety and cost-effectiveness can be clearly assessed.

## Conclusions

Common data elements in clinical trials have been identified and their availability in hospital systems elucidated. Several elements, often those related to reimbursement, are frequently available whereas more specialized elements are ranked at the bottom of the data inventory list. Hospitals that want to obtain the benefits of reusing data for research from their EHR are now able to prioritize their efforts based on this common data element list.

## Additional files

**Additional file 1:** Title of data: CTE & SAE Data Inventory. Description of data: List of common data elements in clinical trials with domain, availability/completeness, occurrence in trials, semantic codes and definition. (XLSX 34 kb)

## Abbreviations

CDE: Common Data Element; CDISC: Clinical Data Interchange Standards Consortium; CRF: Case Report Form; CTE: Clinical Trial Execution; EDC: Electronic Data Capture; EFPIA: European Federation of Pharmaceutical Industries and Associations; EHR: Electronic Health Record; EHR4CR: Electronic Health Record for Clinical Research; IMI: Innovative Medicines Initiative; SAE: Serious Adverse Event; SDTM: Study Data Tabulation Model; SDV: Source Data Verification; TMDb: Trial Master Database; UMLS (Unified Medical Language System)

## Acknowledgments

We would like to thank all Work Package 7 members of the EHR4CR project in their contribution to this research. We would thank especially Andreas Grass, Andy Sykes, Caroline Lafitte, Christel Wouters, Elena Bolaños, Fabien Didier, Florence Botteri, Gunnar Magnusson, Jenny Skogsberg, Louise Scott, Marta García Suárez, Nadir Ammour and Yohann Ndja who were part of the pharmaceutical review group.

We also like to thank all partners at the sites who contributed in the data exports, especially Bolaji Coker, Cezary A. Szmigielski, Colin McCowan, Dina Vishnyakova, James A. Cunningham, Kevin Ross and Sebastian Mate.

## Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant n° 115189, resources of which are composed of financial contribution from European

Union's Seventh Framework Program (FP7/2007-2013) companies' in kind contribution.

## Availability of data and material

The clinical trials' CRFs which were analyzed during the current study were only available to members of the EHR4CR project. The results of the analysis are included in this published article and its supplementary files.

## Authors' contributions

PB designed the study, collected and interpreted the material and wrote the manuscript. MM, EZ, DA, TG and JD obtained the results for their hospital sites and helped to draft the manuscript. MD supervised the methodological approach and supported to draft the manuscript. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

An ethics approval was not necessary for this research. Clinic directors (data owners) were asked for permission to access the data to generate aggregated numbers. This was in line with respect to national data protection laws.

## Author details

<sup>1</sup>Institute of Medical Informatics, University of Münster, Münster 48149, Germany. <sup>2</sup>University of Dundee, Dundee DD2 4BF, UK. <sup>3</sup>Département d'Informatique Hospitalière, AP-HP, Hôpital Européen Georges Pompidou, Paris 75015, France. <sup>4</sup>CHIME, Institute of Health Informatics, University College London, London NW1 2DA, UK. <sup>5</sup>Previously Bayer Healthcare, Building K9, Leverkusen 51368, Germany. <sup>6</sup>Novartis Pharma AG, Basel 4002, Switzerland. <sup>7</sup>Chair of Medical Informatics, University of Erlangen/Nuremberg, Erlangen 91054, Germany.

Received: 22 July 2016 Accepted: 7 November 2016

Published online: 22 November 2016

## References

1. ClinicalTrials.gov Trends, Charts, and Maps. <https://clinicaltrials.gov/ct2/resources/trends>. Accessed: 05 April 2016.
2. Jha AK, DesRoches CM, Kralovec PD, Joshi MS. A progress report on electronic health records in U.S. hospitals. *Health Aff (Millwood)*. 2010;29(10):1951–7.
3. Tipping MD, Forth VE, O'Leary KJ, Malkenson DM, Magill DB, Englert K, Williams MV. Where did the day go?—a time-motion study of hospitalists. *J Hosp Med*. 2010;5(6):323–8.
4. 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.
5. Bruland P, Forster C, Breil B, Ständer S, Dugas M, Fritz F. Does single-source create an added value? Evaluating the impact of introducing x4T into the clinical routine on workflow modifications, data quality and cost-benefit. *Int J Med Inform*. 2014;83(12):915–28.
6. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic PY, Jaulent MC, Daniel C. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. *J Biomed Inform*. 2011;44 Suppl 1:S94–102.
7. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch HU, Ganslandt T. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. *Int J Med Inform*. 2013;82(3):185–92.
8. Liu K, Acharya A, Alai S, Schleyer TK. Using electronic dental record data for research: a data-mapping study. *J Dent Res*. 2013;92(7 Suppl):905–6S.
9. Zahlmann G, Harzendorf N, Schwarz-Boegner U, Paepke S, Schmidt M, Harbeck N, Kiechle M. EHR and EDC Integration in Reality. *Applied Clinical Trials* 2009. <http://www.appliedclinicaltrials.com/ehr-and-edc-integration-reality>.
10. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144–51.

11. Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. *Fam Pract.* 2010;27(1):121–6.
12. Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. *Yearb Med Inform.* 2014;9:215–23.
13. Breil B, Semjonow A, Müller-Tidow C, Fritz F, Dugas M. HIS-based Kaplan-Meier plots—a single source approach for documenting and reusing routine survival information. *BMC Med Inform Decis Mak.* 2011;11:11.
14. Ohmann C, Kuchinke W. Future developments of medical informatics from the viewpoint of networked clinical research. Interoperability and integration. *Methods Inf Med.* 2009;48(1):45–54.
15. Electronic Health Records for Clinical Research. <http://www.ehr4cr.eu>. Accessed: 14 June 2016.
16. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, Karakoyun T, Ohmann C, Lastic PY, Ammour N, Kush R, Dupont D, Cuggia M, Daniel C, Thienpont G, Coorevits P. Using electronic health records for clinical research: the case of the EHR4CR project. *J Biomed Inform.* 2015;53:162–73.
17. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, Sundgren M, Ericson M, Karakoyun T, Coorevits P, Kalra D, De Moor G, Dupont D. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the electronic health records for clinical research (EHR4CR) European project. *Contemp Clin Trials.* 2016;46:85-91. doi:10.1016/j.cct.2015.11.011.
18. Dastgir J, Rutkowski A, Alvarez R, Cossette SA, Yan K, Hoffmann RG, Sewry C, Hayashi YK, Goebel HH, Bonnemann C, Lawlor MW. Common Data Elements for Muscle Biopsy Reporting. *Arch Pathol Lab Med.* 2016;140(1):51-65. doi:10.5858/arpa.2014-0453-OA.
19. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. *Methods Inf Med.* 2006;45(6):594–601.
20. Redeker NS, Anderson R, Bakken S, Corwin E, Docherty S, Dorsey SG, Heitkemper M, McCloskey DJ, Moore S, Pullen C, Rapkin B, Schiffman R, Waldrop-Valverde D, Grady P. Advancing Symptom Science Through Use of Common Data Elements. *J Nurs Scholarsh.* 2015;47(5):379–88.
21. Saver JL, Warach S, Janis S, Odenkirchen J, Becker K, Benavente O, Broderick J, Dromerick AW, Duncan P, Elkind MS, Johnston K, Kidwell CS, Meschia JF, Schwamm L; National Institute of Neurological Disorders and Stroke (NINDS) Stroke Common Data Element Working Group. Standardizing the structure of stroke clinical and epidemiologic research data: the National Institute of Neurological Disorders and Stroke (NINDS) Stroke Common Data Element (CDE) project. *Stroke.* 2012;43(4):967–73.
22. Taruscio D, Mollo E, Gainotti S, Posada de la Paz M, Bianchi F, Vittozzi L. The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration. *Arch. Public Health.* 2014;72(1):35.
23. Doods J, Lafitte C, Ulliac-Sagnes N, Proeve J, Botteri F, Walls R, Sykes A, Dugas M, Fritz F. A European inventory of data elements for patient recruitment. *Stud Health Technol Inform.* 2015;210:506–10.
24. Doods J, Botteri F, Dugas M, Fritz F; EHR4CR WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials.* 2014;15:18.
25. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *AMIA Jt Summits Transl Sci Proc.* 2010;2010:1–5.
26. ICH guidance E6: Good Clinical Practice. Consolidated guideline. US HHS, US FDA, CDER, CBER; 1996. <http://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/GuidancesInformationSheetsandNotices/ucm219488.htm>. Accessed 09 Nov 2016.
27. Talend Open Studio for Data Integration. <http://www.talend.com>. Accessed: 21 January 2016.
28. CDISC Study Data Tabulation Model. <http://www.cdisc.org/standards/foundational/sdtm>. Accessed 2 Feb 2016.
29. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. *Dokl Akad Nauk SSSR.* 1965;163(4):845–8. In Russian. English translation in *Soviet Physics Doklady*, 10(8), pages 707–710.
30. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354–359.
31. Lawrence P. Hanging on the Metaphone. *Computer Language*, Vol. 7, No. 12 (December), 1990.
32. TransCelerate Clinical Data Standards: <http://www.transceleratebiopharmainc.com/initiatives/clinical-data-standards> Accessed: 08 December 2015.
33. Köpcke F, Trinczek B, Majeed RW, Schreiwies B, Wenk J, Leusch T, Ganslandt T, Ohmann C, Bergh B, Röhrig R, Dugas M, Prokosch HU. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Med Inform Decis Mak.* 2013;13:37.
34. Dugas M. Sharing clinical trial data. *Lancet.* 2016;387(10035):2287.
35. Kalra D, Beale T, Heard S. The openEHR Foundation. *Stud Health Technol Inform.* 2005;115:153–73.
36. Clinical Element Mode. <http://www.opencem.org>. Accessed: 06 July 2016.
37. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPN project. *J Biomed Inform.* 2012;45(4):763–71.
38. Christoph J, Griebel L, Leb I, Engel I, Köpcke F, Toddenroth D, Prokosch HU, Laufer J, Marquardt K, Sedlmayr M. Secure Secondary Use of Clinical Data with Cloud-based NLP Services. Towards a Highly Scalable Research Infrastructure. *Methods Inf Med.* 2015;54(3):276–82.
39. van der Lei J. Use and abuse of computer-stored medical records. *Methods Inf Med.* 1991;30(2):79–80.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

