Author's response to reviews

Title: A European inventory of common EHR data elements for clinical trial feasibility

Authors:

Justin Doods (justin.doods@uni-muenster.de)
Martin Dugas (martin.dugas@uni-muenster.de)
Fleur Fritz (fleur.fritz@uni-muenster.de)

Version: 3
Date: 18 November 2013

Author's response to reviews: see over
Dear Sir or Madam,

Hereby I resubmit the revised manuscript "A European inventory of common EHR data elements for clinical trial feasibility" (MS: 2446001331049500) for publication in Trials. Please find the answers to editorial remarks and the reviewer’s comments below.

I am looking forward to receiving your final answer about acceptance.

Sincerely yours,
Justin Doods

Contact:
Justin Doods
Institute of Medical Informatics
Albert-Schweitzer-Campus 1 / A11
48149 Münster, Germany
Justin.Doods@uni-muenster.de
+49 (0)251 – 83 52587

Editorial requests:
1) Please remove the tables from the main body of your manuscript. For more information on where to include the tables please see the following link:
   http://www.trialsjournal.com/authors/instructions/research#preparing-tables.
The tables were prepared and added to the bottom of the document as described in the link.

Reviewer 1: Felix Koepcke

Major Compulsory Revisions
1. If possible remove all references to performing 2 exports and creating 2 data inventories. As nothing appears to have happened between both exports besides adding new data elements to the list (and losing 2 data providers), I cannot see how data inventory 1 was anything but an intermediate version of your results with which you were somehow unsatisfied. Through this, the manuscript would become substantially better to follow. To increase better readability and avoid confusions, all references regarding 2 exports and data inventory versions were removed as they are not substantially necessary for the latest version of the data version and the work that has been done.

2. Describe why you decided on defining your own groups and not use existing categories of Luo 2011 (source 19). Luo’s objective was “To semi-automatically induce semantic categories of eligibility criteria from text and to automatically classify eligibility criteria based on their semantic similarity.” They developed a “generic semantic categorization of eligibility criteria” from 4821 criteria sentences and “measured the prevalence of the categories in 27,278 eligibility criteria from 1578 clinical trials”. In my opinion, the method and coverage of Luo is superior to those of this manuscript (no method for derivation of groups given; data elements from survey and 17 trials) and you should have built on that. You should either change your data groups to the semantic classes of Luo or explain in detail, why Luo’s results for clinical trial criteria in general are invalid for feasibility analysis.
We chose to use our own groups, because our focus was also on EHRs and not only on the protocols. From our point of view, expert-based curation of the data inventory has advantages over an automated, NLP-based approach. This is described in the manuscript more clearly now (first paragraph of the discussion).

3. Diagnosis and procedure are treated very generically ("diagnosis code" and "diagnosis text"), while laboratory findings are considered in detail (41 individual codes / findings). Explain somewhere the rationale behind this decision. Also discuss the meaning of the results for a practical case, which will most likely require specific diagnosis codes. Standard terminologies are used in all sites for diagnoses and procedures, so the retrieval of this information is straightforward. In contrast, laboratory findings are typically less coded in EHRs. This is described in the text now as well (Discussion at the end of the first paragraph).

4. In the discussion, you state that some data providers analysed the availability of data not in the EHR but in some subsystem. Depending on how percentages were calculated and the nature of the subsystem, the results of those data providers might not be comparable with those data providers that used their EHR. Elaborate on this in more detail. The discussion on this aspect was extended by giving an example. See discussion, fourth paragraph.

Minor Essential Revisions
Duplicate nature of methods and results:
5. What was the rationale behind performing one export with survey results and data elements from 5 trials, then extending the list by data elements from 12 additional trials and performing the export again? Was this planned from the start? If not, what led to the desire to extend the list? An iterative approach was chosen as a pragmatic way to see if our method worked and to improve the process step by step. It was planned from the beginning to do data exports in each iteration step to determine the availability of the (new) data elements.

6. Is it wrong to assume that the differences in ranking between the top data elements from export 1 and 2 are due only to 2 sites delivering no results and adding new items? Otherwise I do not see why any changes to the availability of data elements should have happened at the data provider sites in the meantime. The differences between export 1 and 2 regarding the same data elements were only minor. New data elements were added and information about availability of these elements is provided in export 2.

7. There is only one expert group meeting in your figures 1 and 2. Does this mean the decision to extend by 12 studies was made without consulting the expert group? No. Domain experts were involved in selection of the initial and the extended set of studies as well as review of all data elements.

8. @“After each round of exports the results were analysed and a consensus on the data inventory was agreed. For example, one important decision after the first data export was to remove data elements from the inventory that were available in less than half of the source systems. Those data elements were put on a separate list, referred to as 'wish list'. After the second data export, elements were moved to the wish list that were not available or not used at any of the sites.” > how can there be data elements in the second export which are not available in any source system, when all data elements which were available in less than half of all source systems were removed after the first export? After the first iteration the data inventory got extended by data elements from 12 additional studies. The availability of new data elements was evaluated in the second data export. Some of those new elements were not available at the sites, so they were moved to the wish list.
9. Why were percentages grouped for one export and not for the other?  
The data export methodology was refined during the project: The first export was performed to verify if our approach was feasible and it was not clear if all the sites had all required approvals in place to show detailed information, hence percentage groups were used. For the second export all approvals were in place.

10. In the methods section you say that the definition of each data element’s meaning happened after adding data elements from 12 additional studies and therefore after the first export. If so, how reliable are the results from the first export if data elements were not defined at that time?  
See answer to previous question: the first export was performed to test our approach. All elements got validated in the second data export and only those results are now presented in the manuscript.

Methodology lacks detail:
11. Methodology for creating the “initial list” of data elements is missing: Add at least (1) how their importance for pharmaceutical companies was measured, (2) the questions of the survey, (3) who (background/qualification) and (4) how many people were questioned, who and how many responded.  
The pharmaceutical companies were asked to provide data elements they consider "most commonly used" for feasibility by their experience. Ten companies participated, so at least ten people were involved. General information about the background of the partners from EFPIA was added. See methods, last paragraph.

12. How was the heat map created?  
It was created using Excel, described in the text now as well (methods, end of first paragraph).

A brief summary was given (Methods, first paragraph), but as the simplification of eligibility criteria is not the focus of the paper it is not described in detail.

14. How was the expert group composed?  
Background information about the group that did the manual review was added to the methods section, last paragraph.

15. Methodology for selecting the 17 studies is missing.  
This is described in sentence 3 of the material paragraph.

16. Rationale for requiring included trials to “have run at least at one EHR4CR data provider site” and to represent “each EFPIA company” (the EFPIA homepage says there are 73 member companies) is missing.  
The data inventory was created as part of the EHR4CR project involving ten EFPIA companies and therefore all partners were included. This is described in the manuscript now as well. See discussion, paragraph six.

17. Chain of events unclear from methods section without figures. I advise to write in chronological order.  
Methods section was updated accordingly.

18. It remains unclear, what the ranking of data elements is done for. It also appears to make no sense to rank first by percentages and subsequently by availability.  
The ranking of the data exports was done to identify those data elements that are hardly used. From our point of view, ranking by data availability is important because only available data can be re-used. The second part of the remark is answered in point 9.
More details on what the sites did needed:

19. Definition for “availability” should be added.
In the methods section we describe what we mean with availability ("...conducted to capture the availability of each element (available yes/no) and the frequency of documentation ...").

20. Please describe exactly how percentages were calculated.
A description how the relative percentages were calculated was added at the end of the methods section.

21. Reason(s) why 2 sites did not perform a second data export are missing.
Data exports are a laborious manual task and two sites didn’t have the (project) resources available at that time.

22. You mention that some sites used a specialized subsystem instead of the EHR. > Which sites used which systems and what was the reason to have them not use the EHR? > I would like a table with data provider, site ID (if this can be disclosed), description of the analysed IT component (specialty, purpose, ...).
The systems used at the sites depend on the access to the data the local partners have. This is described in the manuscript now. An exact table of which provider used what system and further details cannot be disclosed.

Additional results I would appreciate:

23. To identify common eligibility criteria you performed 1) a survey and 2) an analysis of trial protocols. Can you describe the match between both methods’ results?
The methods are intended to complement one another, a comparison between them was therefore not done.

24. Can anything be said on the frequency of the data elements in your inventory in real trial protocols?
The simplified criteria were derived from real trial protocols. A top list of used eligibility criteria is not subject of this study and therefore we did not measure how often each element was used in our set of studies.

25. Results of meetings / telephone conferences are missing.
Many telephone conferences and meetings were done. We don’t think that exact minutes for each add much to the manuscript, but examples of decisions made are provided. See methods, second paragraph.

26. Results section should describe the results in terms of availability and frequency of data elements with numbers.
The heat map contains the numbers of availability and frequency/usage.

Others

27. The introduction explains the motivation for the EHR4CR project, but not specifically those of this manuscript. It remains unclear who will use the inventory to what purpose.
Our motivation is to increase secondary use of routine hospital data. This was added in the background section (sixth paragraph).

28. “How can a valid and EFPIA accepted inventory be created?” is a research question not being explored by this manuscript. “Creating an inventory” could be an objective.
The research question was slightly reworded, since it was not our intention to investigate different possible processes to create the data inventory.

29. I don’t have the ISO/IRC 11179 to check, but it seems odd to give two definitions for “data element”. Is it possible that the first definition is actually for “data element concept”? The definition on data elements was revised in the manuscript and one of the two references was removed.

30. I found many undefined concepts in data element definition and subsequent paragraph confusing. Definition requires following attributes for each data element: definition, identification, representation, permissible values, value domain, datatype, representation class, unit of measure. Of these the meaning of “definition”, “identification”, “representation”, “value domain”, “datatype” and “representation class” remain undefined. The definition has been made clearer in the manuscript.

31. Furthermore you only define “data group”, “data item” and “data type” for your data elements. In a strict sense, your data elements thus appear to be no data elements according to the given ISO/IRC 11179 definition. Our data elements are compliant with the definition of ISO 11179, but not 100% identical. Because of the different countries, languages and locales involved we didn’t specify the representation part (of ISO 11179) for example.

32. You write in the methods section: “capture the availability of each element (available yes/no)” and “availability and frequency of a data element were captured separately.”. But later: “the availability of the elements (available = 0-100%; not available = N/A)”. The second phrase seems to contradict the first one. The second sentence is a reference to the Top 10 list while the first one was used for the data exports. This should be clearer now with the removal of the Top 10 lists.

33. The analysis seems to be limited to the experience of a given set of pharmaceutical companies and hospitals. Discuss their representativeness for clinical trials in general. How well do the data providers represent the average European hospital and how well do the pharm. companies represent the entirety of clinical trials? The EFPIA partners in the EHR4CR project are among the largest researching pharmaceutical companies in Europe. No references on average European hospitals were found, so no comparison on the representativeness can be made.


Discretionary Revisions
35. The discussion of Weintraub is too detailed. The comparison was revised and shortened.

36. Paragraph Data Inventory mainly repeats paragraph Data Element. This is true, but we wanted to make sure that readers exactly know what we mean with data inventory.
37. First sentence second paragraph discussion repeats first sentence first paragraph discussion. Thank you for the hint, the sentence of the second paragraph was removed.

38. Non-UK source in addition to [1] would be welcome. Preferably something with a European focus. Another source was added from van der Wouden et. al. who surveyed studies conducted in The Netherlands.

39. Is there a source for “If recruitment could be optimized by a better selection of clinical research centers and if better trial protocols could be created through an improved and more accurate feasibility analysis, clinical studies could be completed faster and more cost-efficient.”? We don’t have a published reference for that, but this is what our EFPIA partners conveyed to us during discussions/talks.

40. Same paragraph: Elaborate on how the lack of transparency for the sponsor constitutes a problem for the conduct of clinical trials. Elaborated in the manuscript now.

41. @”but Ross et. al. [3] showed that the majority of criteria in studies are highly complex”. Ross classified 85% of all criteria as complex; A criterion was complex if it included negation (26%), an arithmetic operator (15%) or a Boolean connector (53%). Thus “not pregnant” would already be a complex criterion for Ross, which is quite a low threshold in my opinion. Anyway, Ross uses no category named “highly complex” though one might argue that the 35% of criteria that contained 2 or more semantic patterns could be classified as such. Thank you, this is correct, but we removed this part due to the comment 1.

42. Is there a source for the classification of data as “semi-structured” if it is encoded in a proprietary terminology? This part about local value sets was removed.

43. @”data content and frequency” > maybe something with “element presence” instead of frequency” To our understanding “element presence” would be equivalent to ”availability”, while frequency reflects how often a data element was documented.

44. @”In this process the elements were grouped by their context and the data groups were created as well.” > explain the difference between grouping by context and creating groups Data groups refer to the first part of our data element concepts. Based on the context those ”data groups” where then created (e.g. demographics, diagnosis,…).

45. Give an example of a typical criteria <> core information – pair. An example of how an eligibility criterion was simplified to become a data element is given in the first paragraph of the methods section.

46. In methods you speak of measuring frequency, relative frequency, percentage, relative percentage, numbers and relative numbers. Please unify or explain the difference. The different terms were harmonized.

47. @”There is generally no direct incentive for documentation of time intensive scales …” > From the context it is clear that you mean research related data, but still I fail to understand what time intensive scales are.
The UPDRS is comprised of five parts and part 1 (see http://www.alz.washington.edu/NONMEMBER/UDS/DOCS/VER1_2/b3.pdf) for example consists of 14 questions. Answering all the questions with a patient will take a lot of time.

48. The comparison of the data elements of both versions showed that the top 10 elements are stable although 31 elements got added or removed.” Appears out of context. The sentence was removed as was the whole paragraph about the top 10 elements and its comparison of the two DI versions.

Spelling errors:
49. “where den validated”
50. “The definitions of each element contains”
51. ISO/IEC 11179 Standard defines data element in Part 4 [7] as follows “Unit of data for which the defin ...” > either introduce “:” or remove “follows” Tables and figures: 52. Table 4: decimal places are unnecessary. Thank you for the hints, they were corrected in the manuscript.

53. I suggest making figure 1 and 2 a single figure. The figure was redone as part of removing the references of two iterations.

54. Sum column in heat map should be replaced by average and median. The sum column of the heat map was replaced by avg. usage over all 9 sites.

55. Discharge date should not be coloured in table 4. The table was removed as part of removing all references to two iterations.

56. Quality of the heat map is too low (at least in the materials I received for review). The heat map was included as a pdf now to increase the quality.

57. The heat map seems to be dispensable. It is neither presented nor discussed in the results or the discussion section. I advise to merge it with the supplement table. The heat map is discussed in the discussion section now as well.

58. Similarly, the wish list should be integrated within supplement 1. Following your methods, the wish list should resemble the lower part of the list of all data items, when sorted by availability. We wanted to have the whole data inventory in the manuscript, but due to formatting instructions of the journal it has to be added as a supplement. We think it still makes sense to keep the wish list in the main body, because it shows which elements are frequently used in study protocols, but are not available in EHRs.

59. Mention of figure 1 should be at the beginning of the text. The figure was moved to the beginning of the section.

60. Table 5 and 6 are not necessary. The tables were revised so that they better relate to the data inventory.

Reviewer 2: Betty Tai

Minor Essential Revisions
1. “Background” section:
As defined by ISO/IEC 11179, a data element includes two parts: one is the data concept, and the other is the data representation. The authors assert in the paper (e.g., in the third paragraph of the Background section)—and I would agree—that the reusability of EHR data elements relies largely on the nature of data element representation. However, the current work seems not to focus on the discussion of value domain-relevant issues in European EHR systems, and in fact these issues are considered “out of scope” of the current work (second paragraph of the Discussion section). Instead, in the current paper, the authors describe how to create a data inventory which comprises “data groups” and “data items” that are useful for clinical trials feasibility analyses and that hopefully exist in EHRs. This reviewer, therefore, sincerely hopes that the authors will consider furthering this work to identify data element representation issues in European EHRs systems in the future.

The aspect of value lists is important but also extremely complex, in particular regarding the multilingual documentation of EHRs in different countries. Therefore we consider it out-of-scope for this manuscript.

2. “Methods” section:
The authors summarize the extensive process of developing and vetting the data inventory in Figures 1 and 2. Using figures to describe the process is an excellent way to clearly orient readers. In order to make this illustration more informative, this reviewer suggests the following:

a. Modify the figures by first combining figures 1 and 2 into one.

Figures were updated accordingly.

b. Then, instead of providing a long narrative in the figure legends, add more information in the body of the flow diagram, e.g., # of sites, protocols, percentage groups, etc., as described in the “Methods” section of the main text. This should allow readers to visually appreciate the detailed inventory process much more efficiently

Thank you, we added the information in the figure.

3. “Results” section:
This paper presents an European Inventory of common data elements (a total of 75 data elements initially and then the final 54 data elements) created from extensive and laborious review and selection by experts’ consensus. The data elements identified in this inventory are available in current EHRs used in European medical settings. This inventory will be used for feasibility testing to enable researchers to select study sites and recruit study participants more effectively and efficiently. That would be an incredible achievement. In order to enrich the information in this paper, this reviewer strongly suggests that the authors consider:

a. Adding the “additional file 1” to the body of the paper.

Due to the size of the table and the formatting instructions of the journal the data inventory cannot be added in the manuscript, but was added as an additional file.

b. Reformatting Figure 4 to make it legible. The figure is an effective way to summarize the data elements, but in its present form it is difficult to read.

Figure 4 (now 3) is available in a better quality now.
This reviewer also questions the rationale for ranking the frequency of data elements in the inventory. As I understand it, lower frequencies may not indicate that the EHRs do not include such data elements; it may very well be that the data elements exist, but providers for some reason simply did not collect or record the data. Therefore, the results from an exercise to rank the frequency of data elements in the EHRs, without detailed investigation or analysis, could be misleading.

Data elements were ranked to identify those that are not used in routine care. The data inventory should not only contain elements that are common for research, but also common in routine documentation. When ranking the elements we also identified elements that are available but not used. Some elements were found that are not available at all, so no forms in the EHRs could be identified where those elements can be documented.

Highly ranked data elements could be those data elements that are most common in the EHR systems (e.g., gender, birth date, diagnosis code, etc.). But these most commonly recorded data items may not contain the unique information required for clinical trial eligibility criteria (as these criteria are very disease specific). Even if a disease specific data element of interest happens to have a high frequency, it may simply indicate a high prevalence of this disease in the sites investigated, and this kind of high frequency may be transient due to the site locations, and hospital specialties. Moreover, it does not seem to be very meaningful to lump-sum the frequency of a “common data element” across patients with different diseases from oncology, inflammatory diseases, neuroscience, diabetes, to cardiovascular and respiratory diseases, because data elements on top of the rank list are not likely to be disease specific but are more likely to be common across all disease areas. The risks of this frequency ranking approach are two-fold: it may result in only a small fraction of data elements in the final inventory being useful for a feasibility analysis for any specific disease area, or once the patient population changes, the inventory list of “common data elements” will have to change too.

c. Therefore, this reviewer suggests that the authors also include a discussion of the limitations of the current methodology and perhaps offer potential alternative approach(es) to constructing an inventory, either by disease category or organ system category.

As suggested by the reviewer, a section on limitations of our approach and possible alternatives was added.

4. “Discussion” section:
The authors compare their work with “related work” by others in this section, but this reviewer has doubts about the merit of doing such comparisons for the following reasons:

Comparison to ACCF/AHA 2011’s work (Weintraub, et al): Weintraub’s work had very different goals and objectives from the current work. The ACCF/AHA’s work was intended to develop national professional medical society consensus-based data standards for the EHR vendor community, specifically in the cardiovascular specialty area, to facilitate future data exchange and ensure the interoperability of diverse EHR systems in the United States. Unlike the objectives of the current work, which is directed at clinical trial feasibility analyses (and potentially participant eligibility criteria matching), key elements and definitions in Weintraub’s work were mainly intended for patient management and care quality performance improvement as well as translational research. In the first place, the lack of involving professional medical societies in the review and consensus process for selecting data elements is a limitation of the current effort, compared with ACCF/AHA 2011’s work. Compared with ACCF/AHA 2011’s work, the current project also lacks disease specificity, potentially leading to a scenario in which data elements at the top of the inventory list are
those generically available for most patients, but not uniquely useful for disease specific eligibility criteria matching and feasibility analysis for a specific trial.

The section on "related work" was updated and shortened. We intended to provide a general, not a disease-specific data inventory.

Comparison to Weng’s work: The work by Weng addressed semantic patterns in free-text eligibility criteria and developed a conceptual schema through semantic tree pattern mining and sentence syntactic parsing to generate a semantic network for extracting information from free-text eligibility criteria. The project described in this manuscript does not directly deal with the hierarchical syntax of free text, nor does it deal with semantic relationships between semantic types. The comparison can only indicate that some data items and data groups in the current work are relevant to eligibility criteria matching, which is expected. I would prefer if the authors had compared the current list of 75 (or 54) data elements with the 230 simplified eligibility criteria and data elements identified by the authors’ previous work (reference #9). This latter comparison could potentially shed light on the usefulness of the current ranking list for trial feasibility analysis.

The 12 simplified studies used in #9 (now #8) were also used for this work. #9 and this work are complementary to each other. The comparison table between the data inventory and the semantic classes from Weng/Luo was revised to show how the categories relate to each other.

a. This reviewer suggests that the authors consider eliminating these comparisons.

We believe that these comparisons add value to the paper. Weintraub et. al. was shortened and Weng et. al. was reformulated.

5. Other revisions

a) If the authors wish to keep Table 5, it needs be modified. The title of column 1 should not be “Overview of corresponding data fields and data elements.” The table should also present the total number of data elements (for current work)/data fields (for AHA 2011 work) in each corresponding “data group” of the current work.

The table was updated but an extension of the table was not done, because it would be hard to read without adding much additional information for the reader.

b) In the second paragraph of the “Methods” section, the numerical ranges for each “percentage group” should be mutually exclusive.

This was updated and the part of the manuscript was moved to the discussion (third paragraph).

c) In the “Lessons learned” section, one redundant “plausibility” should be deleted from this list: “(completeness, correctness, plausibility, concordance, plausibility and currency).”

d) In the “Results” section of the abstract, the final sentence should be revised to make it clear that the sub-list (“wish list”) is comprised of the 21 elements that were removed from the data inventory.

Thank you for these comments, they were changed accordingly.

Reviewer 3: Eugenio Santoro

no comments
Reviewer 4: Rachel L Richardson

MAJOR COMPULSORY REVISIONS:

This is a nicely written paper on an important topic, and this reviewer applauds authors for clearly describing the importance of common data elements to support care delivery and research. Further, authors have done a nice job identifying an approach to this problem in the form of an inventory of data elements that is derived both from elements available in EHRs and important to research. However, I think that the investigation authors have done here is only preliminary and not ready for readers hoping to identify and implementing standards. Authors have not provided enough detail on the methods for readers to assess if their collection of data elements is relevant (either from the perspective of HER developers/implementers or researchers). For example, it is not clear how many EHRs were surveyed or the details about them (e.g., type of system, type of facility, nature of patient populations, size, etc.).

The number of surveyed EHRs was added. Type of system and facility were described very generally. The data inventory is based on input from eleven hospitals in five countries as well as ten major pharma companies, therefore it is more than only preliminary from our point of view.

In addition, there are no details about who you surveyed in the pharma companies – how many people were surveyed as part of your consensus process and what was their background and training?

The number of people who participated in the consensus was added. Background of the pharmaceutical specialists of the group who manually checked each data element was added.

What types of protocols and studies and forms did you look at for research elements? What disease areas?

A table with disease areas and the number of studies was added as well as more information about the study types. See material, first paragraph.

The approach I would suggest is to identify clearly important data elements for research (perhaps by surveying a set of CRFs or protocol eligibility criteria as you did but in a bigger way with more details on the sources of your research data elements and supporting data (e.g., type of study, medical domain, type of form, nature of research questions) to demonstrate the representativeness or generalizability of these elements to other research domains. Then, I would attempt to match these elements to data availability in various EHRs and report the coverage of elements (and likely the different variations of value sets and definitions) that are available in different EHRs.

The number of data elements might increase with more iterations. However, the differences between the first and second iteration of the data inventory were relatively small, so we consider the current version as representative. The heat map shows the coverage and availability of the data elements of various EHRs.

Although, overall the paper is logically organized and nicely written, it would benefit from some editing. The description of multiple iterations of the inventory are confusing and could be condensed to describe the (iterative) development of the final inventory.

The description of inventory iterations was updated.

The term ‘feasibility analysis (2nd paragraph Background section) needs to be defined more directly and subsequently labeled as ‘trail’ or ‘site’ feasibility.

This is now described as ‘site feasibility analysis’.

This could be an important and central theme throughout the paper. Overall, paper is fairly abstract and grounded examples would help.

More examples were added to the manuscript, e.g. "Patient with confirmed deep vein thrombosis’ was for example simplified to ‘Diagnosis/Text: deep vein thrombosis’ in the methods section.
Good point about complexity of eligibility criteria in background. Although it provides nice context, it is distracting to your study and should be removed.
The part about complexity of criteria was removed from the manuscript.

More relevant would be text on other CDE elements, including NCI, LOINC, USHIK, CDISC, and FDA efforts – and probably many others. Similarly the details on data element definitions might be too much for this paper.

CDE elements in general are a very broad topic and beyond the scope of this manuscript. Our main topic is a data inventory to support site feasibility analysis in clinical research.

Other editorial suggestions include:
Streamlining the Abstract to be a bit more concise.
We shortened the abstract as suggested.

Key points to focus on are
study planning (perhaps focus on site identification) and clearly defining your methods – including details regarding terms like “consensus-driven approach.” For example, the abstract background could be trimmed and focused as below. Clinical studies are a necessity for new medications and therapies, but many fail to meet their recruitment numbers. Study planning could be supported by more reliable site identification in the feasibility analysis. The increasing number of electronic health records (EHRs) in hospitals creates huge databanks emerge that could be utilized to support research. The Innovative Medicine Initiative (IMI) funded project “Electronic Health Records for Clinical Research” (EHR4CR) Has developed a standardized and homogenous inventory of data elements to support utilization of EHRs for research. We developed an Inventory of elements for feasibility analysis of new trials.

The abstract was updated as suggested.

Further the examples that authors provide are not well modeled data elements nor should they be promoted as standards of any kind. For example, in the attachment (appx), the first data element DOB does not specific format at MM/DD/YYYY or DD/MM/YYYY and so even in the example (10/10/1967) it is not clear. The description of lab data elements should be more clear and reference LOINC and CDIDSC and HL7 models that are fairly well developed and standardized.

We revised the definition of data elements to be compliant with ISO 11179. No specific formats /representations were specified however, as they may vary, depending on the national locales and systems used. For most of the Lab values LOINC codes were used to define the context of the data element.