SOFTWARE TOOL ARTICLE

The iCRF Generator: Generating interoperable electronic case report forms using online codebooks [version 1; peer review: 2 approved with reservations]

Sander de Ridder, Jeroen A.M. Beliën

Department of Pathology, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, 1081 HV, The Netherlands

Abstract
Semantic interoperability of clinical data is essential to preserve its meaning and intent when the data is exchanged, re-used or integrated with other data. Achieving semantic operability requires the use of a communication standard, such as HL7, as well as (functional) information standards. Manually mapping clinical data to a medical thesaurus, such as SNOMED CT, is complicated and requires expert knowledge of both the dataset, including its context, and the thesaurus. As an alternative, the (re-)use of codebooks, data definitions which may already have been mapped to a thesaurus, can be a viable approach.

We’ve developed the iCRF Generator, a Java program that can generate the core of an interoperable electronic case report form (iCRF) for several of the major electronic data capture systems (EDCs). To build their CRFs, users can select one or more items from established codebooks, available from an online system called ART-DECOR. By providing an easy to use method to create CRFs for multiple EDCs based on the same codebooks, interoperability can be more easily attained.

Keywords
Interoperability, eCRF, iCRF, Codebook, FAIR, Software, EDC, Clinical data

Corresponding authors: Sander de Ridder (s.deridder1@amsterdamumc.nl), Jeroen A.M. Beliën (jam.beliën@amsterdamumc.nl)

Author roles: de Ridder S: Conceptualization, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; Beliën JAM: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was financially supported by the Dutch Cancer Society (KWF) TraIT2HealthRI (grant 8166). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 de Ridder S and Beliën JAM. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: de Ridder S and Beliën JAM. The iCRF Generator: Generating interoperable electronic case report forms using online codebooks [version 1; peer review: 2 approved with reservations] F1000Research 2020, 9:81 (https://doi.org/10.12688/f1000research.21576.1)

Introduction
Clinical data is essential for health research. Traditionally, such data was captured using paper case report forms (CRFs) and entered into a database manually. Nowadays, the data is often captured directly with electronic CRFs (eCRFs) in an electronic data capture (EDC) system. This has improved the quality of the captured data as well as decreased costs for data collection (e.g., 1, 2).

To allow the captured data to be used beyond its original purpose requires the data to be FAIR (Findable, Accessible, Interoperable and Reusable)³. By making data semantically interoperable, it can be exchanged between systems whilst preserving the meaning of the data⁴. Furthermore, it allows multiple data sources to be combined and understood by computers, thereby e.g. facilitating clinical decision support systems⁵. Hence, when setting up a new data collection protocol, the eCRF should be designed with interoperability in mind. Achieving semantic interoperability requires the use of a communication standard, such as HL7, as well as (functional) information standards⁶, such as the NCI thesaurus or SNOMED CT. However, mapping study-specific terminology to a thesaurus requires expert knowledge of the thesaurus, the data and its context. Therefore, reusing existing codebooks from studies and well-known datasets or CRF templates, such as available from CDISC’s CDASH, the Meta Data Models website and the University of Wisconsin-Madison, can be a viable alternative. Reusing these elements at the very minimum facilitates interoperability with other datasets using these definitions. Furthermore, in many cases well-known codebooks have already been mapped to a thesaurus. For example, in the Basic Health Data Set, which is the standard that will be used by hospitals to exchange healthcare data in the Netherlands (available here, Dutch only), many of the items have been mapped to SNOMED CT.

In this paper, we introduce the iCRF Generator, a program that allows users to easily generate interoperable electronic case report forms (iCRFs) based on online codebooks, thereby improving the interoperability of clinical data collected in and between EDCs. Whereas normally CRF generation is an integrated part of the EDC (e.g. Castor EDC, REDCap), our program can generate the core of a CRF for multiple EDCs. At this time, three systems are supported: Castor, OpenClinica 3 and REDCap. The program allows a user to select one or more codebooks available from an online system called ART-DECOR which allows, amongst others, the storage of dataset definitions and select items of interest, including their codelists. The program currently supports six codebooks, which are further described in the Methods section.

Methods
Implementation
The iCRF Generator was written in Java 8 and later migrated to Java 12 for JavaFX compatibility. Dependencies are managed using Maven and include: JavaFX and ControlsFX for the UI, Apache POI for Excel file management and Log4j for logging. A ZIP file of the iCRF Generator distribution is available for both Mac and Windows. It includes a Java Runtime Environment to ensure independence of the locally installed Java version and ensures the program works out of the box. Source and distribution files are available on GitHub: https://github.com/aderidder/iCRFGenerator/.

Supported codebooks
The iCRF Generator is designed to use codebooks defined in ART-DECOR. ART-DECOR is an open-source tool suite that supports the creation and maintenance of HL7 templates and allows the storage of dataset definitions. Nictiz, the centre of expertise for eHealth and the Dutch SNOMED-CT release centre, facilitates ART-DECOR to create health information standards that are publicly accessible. The iCRF Generator currently offers access to six of these codebooks, which were chosen because of their national relevance (codebooks 1, 2 and 3) and our involvement (4, 5 and 6):

1. The Clinical Building Blocks (Zorginformatiebouwstenen): information models of minimal clinical concepts. They are used as the basis for the Basic Health Data Set.
2. The Basic Health Data Set (Basisgegevensset Zorg): codebook used for the standardised exchange of patient data between e.g. healthcare providers. Implementation of this set is prioritised in healthcare systems like electronic health records. The Basic Health Data Set is aligned with the European Patient Summary.
4. Cancer Core Europe: a European cancer research alliance which aims at bringing together the expertise and critical mass necessary to make translational research available in the clinic.
5. The PALGA Colon biopsy protocol: PALGA is the nationwide network and registry of histo- and cytopathology in the Netherlands.
6. The PALGA Colorectum carcinoma protocol.

Some of the codebooks are available in English, as well as Dutch. Other codebooks may be made available in the iCRF Generator in the future when they become available in ART-DECOR provided they are of sufficient quality and are complementary.

Operation
A standard PC or Mac should be able to run the program without any issues. The program was tested on a Windows 7 PC, a Windows 10 PC and on a virtual machine with OS X El Capitan. To give an indication of the program’s memory usage: selecting a single codebook, the program uses around 230 megabytes of memory; increasing this number to eight codebooks increased the memory usage to 420 megabytes. An internet connection is required, as the program retrieves metadata as well as codebooks from ART-DECOR via the REST API.
Use cases
A typical use case for the iCRF Generator is in the design phase of a study or registry. When a decision has been made on what clinical data is going to be collected, the data manager has to design and build the CRFs for data collection. Instead of doing this manually, the iCRF Generator can be used to select items from the available codebooks and generate the basis of the case report forms. Figure 1 illustrates the iCRF Generator’s complete workflow. In a typical use case, the user first selects an EDC from the dropdown (Figure 2). The user then clicks the “Run” button, after which the wizard interface is started. The first wizard page asks the user to select one or more codebooks (Figure 3). When a user has selected a codebook and presses the “Next” button, a REST-call is made to retrieve an XML which contains which versions and languages are available for the codebook. The second page shows this information to the user, allowing selection of the versions and languages of interest (Figure 4). When the user proceeds to the next page, the selected codebooks are retrieved via one or more REST-calls. These XML files are parsed and the information it contains about the items and their possible values is shown on the third page (Figure 5). This EDC-specific page allows users to select and customise the items that have to be included in the CRF. The final page of the wizard shows a short summary of the number of selected items. Upon completion of the wizard, the program generates the CRF in the format required by the selected EDC and the file is saved to disk. This file can then be imported into the EDC-system or opened in an editor of choice.

Discussion
To allow for the use of data beyond its original purpose, it is essential that the data is FAIR (Findable, Accessible, Interoperable and Reusable)\(^1\). To preserve meaning and intent of clinical data when it is exchanged, requires the data to be semantically interoperable, which requires the use of content standards. Manually mapping data definitions to a medical thesaurus such as SNOMED CT is complicated and time consuming. Instead, reusing existing published definitions (codebooks) can be a viable alternative, as interoperability with other datasets using these codebooks is usually easily achieved. Furthermore, the codebooks may already have been mapped to a thesaurus.

In this paper we introduced the iCRF Generator, a program which can generate electronic case report forms for three major electronic data capture systems: Castor, REDCap and OpenClinica 3. The program allows a user to select items and codelists from several highly relevant codebooks available from the online system ART-DECOR. By providing an easy to use program to generate CRFs using these codebooks, data will be collected using the same definitions, which enhances the interoperability and FAIRness of the data.

Local caching of codebooks
One important usability aspect of software is the software’s performance. The codebooks available in the iCRF Generator are parsed from XML files generated by ART-DECOR. It takes ART-DECOR around 30 seconds to generate the XML file
Figure 2. User selects an electronic data capture system (EDC) from the dropdown.

Figure 3. User selects one or more codebooks.
Figure 4. User selects codebook version(s) and language(s).

Figure 5. User selects items on the left-hand side in the tree and customises an item’s details in the electronic data capture system (EDC)-specific right-hand side.
for the National Institute for Public Health and the Environment screening codebook. Furthermore, in some cases XML files can reference other XML files, which then have to be downloaded and parsed as well. If a user has to wait every time a codebook is selected, user acceptance will quickly erode. Hence, we introduced local caching of downloaded codebooks, which makes usage nearly instantaneous once the codebook is locally available. Furthermore, we intend to make a ZIP file of the cache available for download. Note that downloading a codebook XML from ART-DECOR only takes place if it is accessed for the first time or when a new version of a codebook becomes available and is selected by the user.

Additional EDCs
The iCRF Generator can easily be expanded to include additional EDCs, such as OpenClinica 4, Research Manager and Alea if there is demand and the import formats are available. If desirable, additional internationally established formats, such as CDISCODM, may also be included in the future.

Additional codebooks
At this point, the iCRF Generator gives access to six codebooks, some of which support multiple languages and multiple versions. When more codebooks are published in ART-DECOR, we intend to make these available in the program, provided they are of sufficient quality and are complementary to the codebooks already available. Criteria to decide whether a codebook will be made available will be established in the near future. We believe that providing too many overlapping codebooks will be counterproductive for interoperability as items may well map to different thesauri. Furthermore, the user may be overwhelmed by the available information.

EDC-specific item customisation within iCRF Generator
When an item is selected in the item tree, a user can customise the item. As each EDC has different requirements for its CRFs, the customisation options we provide vary per EDC. As an example, OpenClinica 3 has a “Field Type” (e.g. “Radio”, “Single-Select”) and a “Data Type” (e.g. “ST”, “INT”), whereas in Castor the data type does not exist as a separate entity.

The customisation options we currently provide are limited. The iCRF Generator’s purpose is to facilitate generation of interoperable CRFs. Hence, if everything could be customised, for example replacing the codes in coders with custom codes, it would undermine the purpose of the program. Furthermore, the iCRF Generator is work-in-progress and some further item customisation may be added in the future.

Similar work - alternative solutions & templates
A tool somewhat similar to our own is ODMedit. ODMedit provides a web-based interface to allow users to create a CRF based on elements stored in the Meta Data Repository. When a user has finished creating the CRF, it can either be downloaded in ODM format or uploaded to the Medical Data Models-portal. From there it can be downloaded in multiple formats.

ODMedit differs from our software in several ways. Whereas ODMedit immediately provides access to all items in its repository, we keep our items grouped by codebook. Furthermore, with ODMedit users can immediately add new and edit existing items, and new items are automatically made available in the repository. In our tool we are providing access to only handpicked codebooks, from which the user can select items and customisation of these items is kept to a bare minimum. By allowing users to select items from well-known and supported codebooks only, we believe it should be easier to find the correct item - e.g. if you need pathology definitions, use items from the pathology codebooks. However, we may have to add a search function at some point to make it easier to find items within a codebook. Another difference is that we decided to explicitly ask for which EDC tool the user wishes to create the CRF to allow for EDC specific options. On the other hand, ODMedit does support some features which we do not yet support, such as a repeating group. We may add this at a future time.

An OpenClinica 3 specific CRF generator is also available. This tool converts a csv file to Excel and provides a user with an interface to edit the CRF. However, the tool does not facilitate interoperability.

Multiple initiatives exist that aim at providing templates to improve interoperability. We list several such initiatives below. The National Institute of Health offers Common Data Elements, data elements that are common to multiple data sets across different studies. CDASH, provided by CDISC, gives guidance for developing CRFs used in clinical trials. The OpenClinica Building Blocks developed by TraIT provide OpenClinica users with templates to which they can add study-specific items and remove items that are not necessary for their study. The Australian Government launched a platform for digital health. They provide an extensive library of documents, tools and much more for implementers and developers. The Global Alliance for Genomics & Health (GA4GH) has several workstreams, amongst which one for Clinical & Phenotypic Data Capture, that “Supports the clinical adoption of genomics throughestablishing standard ontologies and information models to describe the clinical phenotype for use in genomic medicine and research, including the capture and exchange of information between electronic clinical systemsand research.”

Software availability
Source code available here: https://github.com/aderidder/iCRFGenerator

Licensed: GNU GPL v3 license.

Acknowledgements
We thank Jan-Willem Boiten (Lygature), Gerben Rienk Visser (Trial Data Solutions) and Maarten Ligtvoet (Nictiz) for reviewing this paper and providing invaluable suggestions. We thank Wessel Sloof (UMCG) for testing the generated REDCap exports.
PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

Reference Source

PubMed Abstract | Publisher Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 14 February 2020

https://doi.org/10.5256/f1000research.23777.r59545

© 2020 Dugas M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Martin Dugas 🔗
Institute of Medical Informatics, University of Münster, Münster, Germany

The iCRF Generator: Generating interoperable electronic case report forms using online codebooks

This manuscript is about iCRF generator, a software tool to generate interoperable CRFs. To foster re-use of data and implement FAIR principles are important topics. Source code of iCRF-generator is available, which is a plus.

Major comments

page 2, Methods, Supported Codebooks: Quantitative Information on 6 codebooks is missing: how many item definitions per codebook are available? how many codelist(items)?

page 2, Introduction: iCRF generates CRFs in Castor, OpenClinica3 and REDCap-Format. Why not CDISC ODM, which is endorsed by FDA? The Portal of Medical Data Models provides CRFs in 18 formats

page 3, Discussion “Manually mapping data definitions to a medical thesaurus such as SNOMED CT is complicated and time consuming”
  => Yes, this is absolutely correct. But it should be mentioned, that this mapping was already done for large sets of data items. For example, the portal of medical data models provides ~520,000 data items with manual (physician-based) terminology mappings. The portal of medical data models is not just a website, it is a registered European information infrastructure.

page 6, Additional EDCs: CDISC ODM is not just an internationally established format, it is required by regulatory authorities (FDA). Define XML (FDA) is based on CDISC ODM!

page 6, EDC-specific item customisation: processing of individual data items is a very work-intensive process, given the high number of data items in clinical studies (~ 500 - 2000). Re-Use of itemgroups (sets of data items) would be useful (this feature is available in ODMedit, which is covered in the
discussion)

page 6, Discussion, similar work: It should be mentioned, that ODMedit provides semantic coding (especially UMLS codes) for data items and codelists. Is semantic coding available in iCRF?

Minor comments
page 2, Introduction, 2nd paragraph:
HL7 is not a communication standard, but a standards developing organisation

page 2, Introduction, 2nd paragraph: the "Meta Data Models website" correct term: Portal of Medical Data Models

page 2, Introduction, 2nd paragraph: Link to Basic Health Data Set (available contents!) should be updated

Figure 2 & Figure 4: should be shrunked, a lot of blank space

References

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Medical Informatics, Medical Data Models, semantic annotations, interoperability

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
Overall
The authors' idea of creating an iCRF generator could alleviate pressure from data managers by making CRFs interoperable.

Introduction
The authors have presented all relevant concepts needed to understand the rest of the article.

Methods
- **Codebooks**: Only 2 of 6 currently supported codebooks are generic enough to be used by users from a wide range of fields. Because the other four are all related to cancer, this could exclude a large group of users.

- **Creating the CRFs**: One of the arguments made in the article is that automating CRF will help people make CRFs more easily. However it is unclear whether using the CRF generator would actually save time. In fact, the figures imply that using the iCRF generator requires a lot of manual work as well.

**Figure 1**
- In this figure, it is at first not clear where to start. We recommend using numbering to clarify the sequential steps.

- It is unclear from the figure and text how the workflow is structured. In fact, the figure seems to imply that there are three different workflows depending on the EDC that is selected by the user. This could be a potential hazard in terms of sustainability as adding an EDC would be a lot of work. A generic workflow with EDC-specific export functionalities would make the program much more future-proof.

- What does the summary show?

**Figure 3**
To help users, we suggest to implement a way of providing information about the codebooks and their content. For example, a link to more information about each codebook or a dropdown view that shows the items in each codebook. This is not there currently and would require users to start searching for it themselves if they're not familiar (enough) with the content of the codebooks.
Use cases
We would like to see a clearer example that also shows what each step in figure 1 looks like.

Discussion
- The discussion mentions that codebooks might be added when of sufficient quality. How do you define sufficient quality and who decides that?

- The authors mention that ODMedit's flexibility is a weakness compared to the iCRF generator. We believe that this could in fact be a strength if introduced in the iCRF generator. In fact, allowing users to contribute to the codebooks and item lists would widen the user base. To avoid the creation of faulty or multiple items denoting the same term, the authors could consider allowing users to contribute, but establish a curated group of codebooks and elements. This makes the software tool more flexible for its users.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: Petros Kalendralis (co-reviewer) works on the Trait2HealthRI research project

Reviewer Expertise: Our area of research focuses on making medical data FAIR. For example, we employ the Personal Health Train to share information across medical centers in a privacy-preserving manner.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com