SOFTWARE TOOL ARTICLE

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

Sander de Ridder, Jeroen A.M. Beliën

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

v2 **First published:** 04 Feb 2020, **9**:81
https://doi.org/10.12688/f1000research.21576.1
Latest published: 23 Mar 2020, **9**:81
https://doi.org/10.12688/f1000research.21576.2

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

Open Peer Review

Reviewer Status ✔ ? ✔

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 23 Mar 2020	✔ report		✔ report
version 1 04 Feb 2020	? report	? report	

1 **Rianne Fijten**, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands
Petros Kalendralis, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands

2 **Martin Dugas**, University of Münster, Münster, Germany

3 **Hugo Leroux**, Australian e-Health Research Centre, Brisbane, Australia

Any reports and responses or comments on the article can be found at the end of the article.

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

How to cite this article: de Ridder S and Beliën JAM. **The iCRF Generator: Generating interoperable electronic case report forms using online codebooks [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2020, **9**:81
https://doi.org/10.12688/f1000research.21576.2

First published: 04 Feb 2020, **9**:81 https://doi.org/10.12688/f1000research.21576.1

REVISED **Amendments from Version 1**

This new version contains revisions made to address the concerns of the reviewers:

Introduction

* Updated the HL7 URL
* "Meta Data Models" corrected to "Portal of Medical Models"
* Added URL to Basic Health Data Set
* Usefulness of iCRF Generator stated more clearly for reusing codebooks which are already mapped to medical thesauri

Methods

* Number of items and codelists in codebooks added

Use cases

* Usefulness of iCRF Generator stated more clearly for reusing codebooks which are already mapped to medical thesauri
* Linked Figure 1 more clearly with Figure 2–Figure 6.
* Added Figure 6, screenshot of summary with label
* Added numbers to steps in Figure 1.
* Decreased blank space in Figure 2 and Figure 3 to decrease figure size

Discussion

* Updated the text to stress that mappings already exist for a large number of items.
* Updated the text with the remark that support for CDISC ODM is now on the roadmap
* Rewrote "Additional codebooks" section to more clearly indicate which codebooks are of interest to us and how the number of codebooks could be improved
* Added a paragraph to the "EDC-specific item customisation within iCRF Generator"

Any further responses from the reviewers can be found at the end of the article

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 data[4]. Furthermore, it allows multiple data sources to be combined and understood by computers, thereby e.g. facilitating clinical decision support systems[5]. 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 standards[4], 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 Portal of Medical 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. Hence, if a codebook is mapped to a medical thesaurus, the iCRF Generator allows the user to use these mappings, preventing the labour-intensive manual mapping. 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). The number of items mentioned below are estimates, as codebooks in ART-DECOR may inherit items from other codebooks multiple times.

1. The Clinical Building Blocks (Zorginformatiebouwstenen): information models of minimal clinical concepts. They are used as the basis for the Basic Health Data Set. The 2017 set contains 100 building blocks, with about 940 items and 211 codelists.

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. The codebook (version 2017) contains 3742 items, of which 876 have codelists.

3. The National Institute for Public Health and the Environment's national screening codebook of bowel cancer and cervical cancer (RIVM bevolkingsonderzoeken). This codebook (version 2019) contains 258 items, of which 103 have codelists.

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. This codebook (version 2017) contains 104 items, of which 93 have codelists.

5. The PALGA Colon biopsy protocol: PALGA is the nationwide network and registry of histo- and cytopathology in the Netherlands. This codebook (version 33) contains 70 items, of which 45 have codelists.

6. The PALGA Colorectum carcinoma protocol. This codebook (version 59) contains 198 items, of which 121 have codelists.

Some of the codebooks are available in English, as well as Dutch.

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, a virtual machine with OS X El Capitan, a iMac and a MacBook Pro both running OS Catalina 10.15.3. 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 manually designing the items and mapping them to a medical thesaurus, which takes a lot of time, the iCRF Generator can be used to select items which have already been mapped from the available codebooks and generate the basis of the case report forms. Figure 1 illustrates the iCRF Generator's complete workflow and Figure 2–Figure 6 show actual examples of the workflow. In a typical use case, the user first selects an EDC from the dropdown (Figure 1, step 1; 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 1, step 2; 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 selected

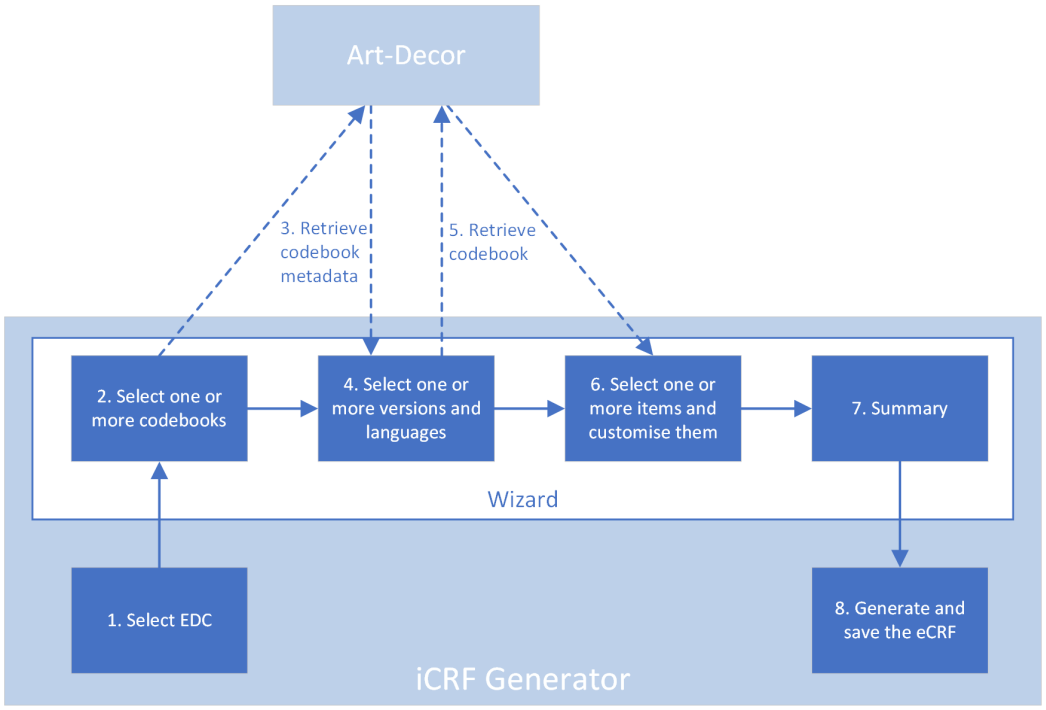

Figure 1. Flowchart showing steps involved in using the iCRF Generator. Dotted lines are calls made by the program to ART-DECOR's REST services.

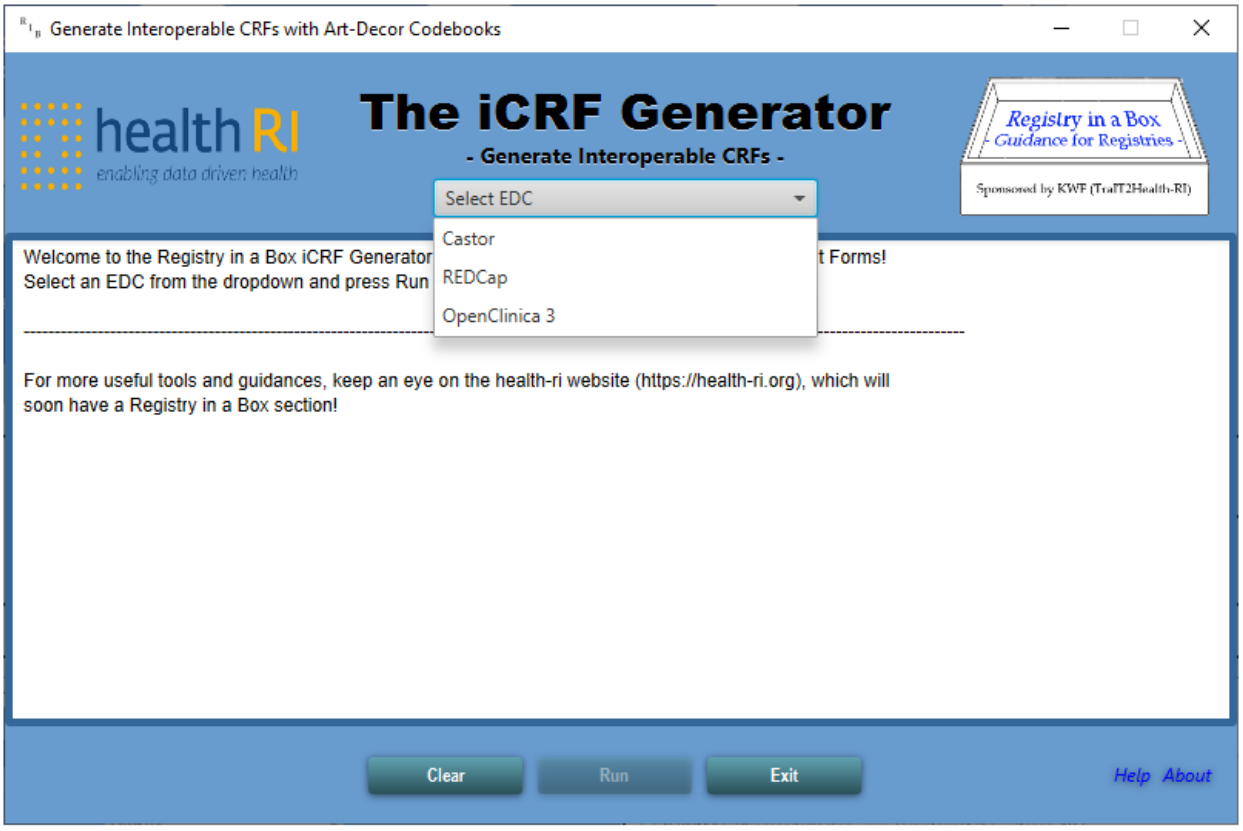

Figure 2. User selects an electronic data capture system (EDC) from the dropdown.

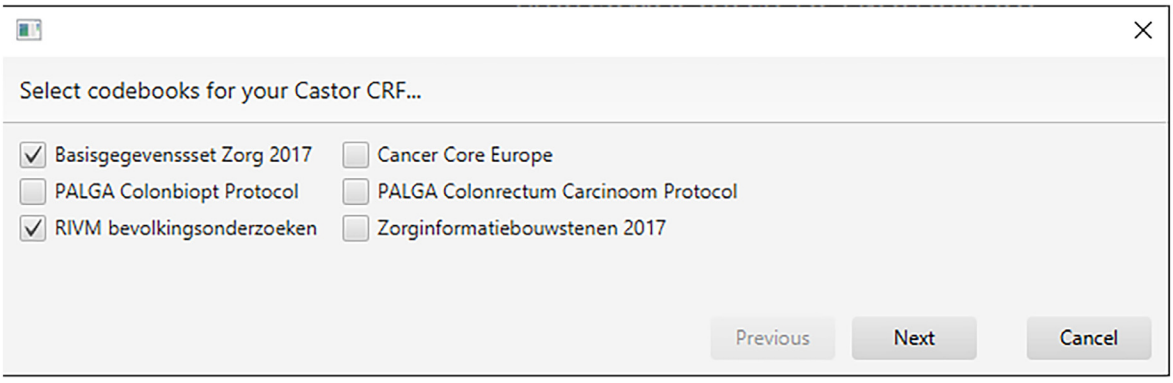

Figure 3. User selects one or more codebooks.

codebooks (Figure 1, step 3). The second page shows this information to the user, allowing selection of the versions and languages of interest (Figure 1, step 4; Figure 4). When the user proceeds to the next page, the selected codebooks are retrieved via one or more REST-calls (Figure 1, step 5). These XML files are parsed and the information it contains about the items and their possible values is shown on the third page (Figure 1, step 6; 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 (Figure 1, step 7; Figure 6). Upon completion of the wizard, the program generates the CRF in

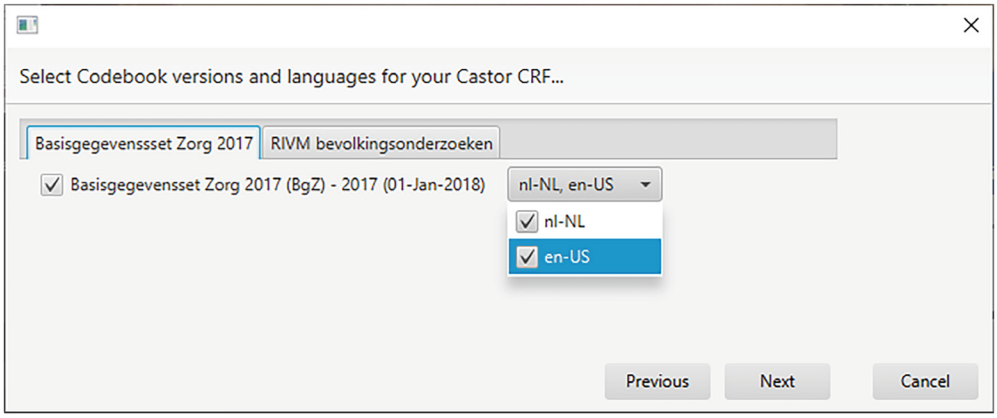

Figure 4. User selects codebook version(s) and language(s).

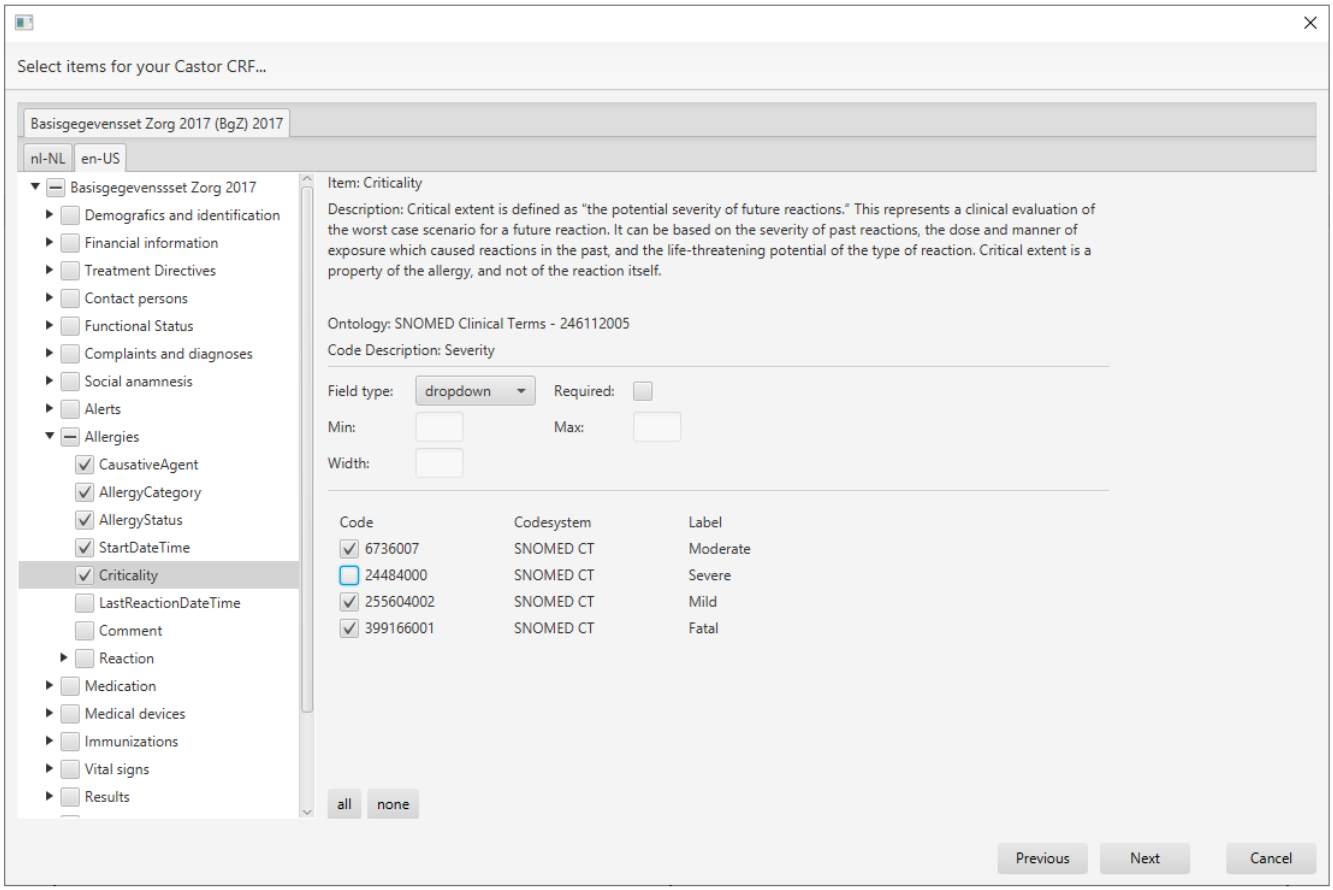

Figure 5. User selects items on the left-hand side in the tree and customises an item's details in the electronic data capture system (EDC)-specific right-hand side.

the format required by the selected EDC and the file is saved to disk (Figure 1, step 8). 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)[3]. 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. For many items, however, mappings are already available (e.g. in ART-DECOR, the Portal of Medical Data

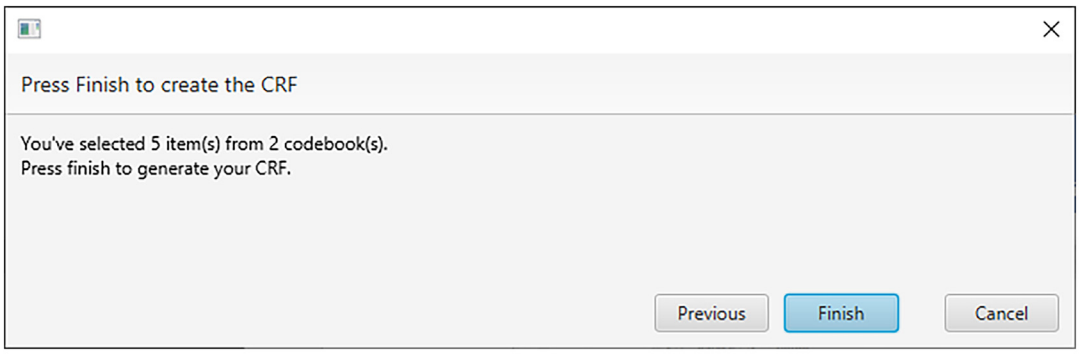

Figure 6. Summary shows number of selected items.

Models and CDISC's CDASH). Reusing these definitions (codebooks) can, therefore, be a viable alternative, as interoperability with other datasets using these codebooks is usually easily achieved.

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 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. Support for CDISC ODM is on our roadmap. If desirable, additional internationally established formats, may also be included in the future.

Additional codebooks

At this point, the iCRF Generator gives access to six nationally established codebooks, some of which support multiple languages and multiple versions. To improve the user-base for the iCRF Generator, the number and variety of codebooks available must increase. While nationwide standards, such as the Basic Health Dataset, can be readily made available, some form of governance may have to be put in place to establish other types of curated codebooks. This could stimulate the community to help build high-quality codebooks, mapped to medical thesauri, while preventing too much codebook redundancy and conflicting items. Furthermore, internationally established codebooks, such as CDISC's CDASH can also be made available.

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.

By adding this EDC-specific customisation, the iCRF Generator's code is more difficult to maintain. However, by allowing the data manager to customise essential fields, the iCRF Generator can provide a ready-to-use CRF. This enhances the user experience, making it a worthwhile investment. 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 codelists 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[6]. 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[7]. 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 through establishing 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 systems and research."

Software availability

Source code available here: https://github.com/aderidder/iCRF-Generator

Archived source code as at time of publication: https://doi.org/10.5281/zenodo.3563500[8]

License: 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 also thank Wessel Sloof (UMCG) for testing the generated REDCap exports.

References

1. Sahoo U, Bhatt A: **Electronic data capture (EDC)--a new mantra for clinical trials.** *Qual Assur.* 2003; **10**(3–4): 117–21.
PubMed Abstract | Publisher Full Text

2. Thriemer K, Ley B, Ame SM, *et al.*: **Replacing paper data collection forms with electronic data entry in the field: findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar.** *BMC Res Notes.* 2012; **5**(1): 113.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Ashrafi N, Kuilboer JP, Stull T: **Semantic Interoperability in Healthcare: Challenges and Roadblocks.** In *STPIS@ CAiSE.* 2018; 119–122.
Reference Source

5. Ahmadian L, van Engen-Verheul M, Bakhshi-Raiez F, *et al.*: **The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey.** *Int J Med Inform.* 2011; **80**(2): 81–93.
PubMed Abstract | Publisher Full Text

6. Dugas M, Meidt A, Neuhaus P, *et al.*: **ODMedit: uniform semantic annotation for data integration in medicine based on a public metadata repository.** *BMC Med Res Methodol.* 2016; **16**(1): 65.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Gaddale JR: **Clinical Data Acquisition Standards Harmonization importance and benefits in clinical data management.** *Perspect Clin Res.* 2015; **6**(4): 179–83.
PubMed Abstract | Publisher Full Text | Free Full Text

8. de Ridder S: **aderidder/iCRFGenerator: First release (Version V1.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3563500

Open Peer Review

Current Peer Review Status: ✔ ? ✔

Version 2

Reviewer Report 11 June 2020

https://doi.org/10.5256/f1000research.25280.r63637

✔ **Hugo Leroux**
Australian e-Health Research Centre, Brisbane, Qld, Australia

This paper describes a iCRF generator that is able to generate codebooks that can be used in three EDCs, namely CASTOR, OpenClinica and REDCap.

This is a very important and useful tool for the clinical research community in promoting and achieving semantic interoperability and the authors should be commended on their effort.
While the tool offers the ability to map the fields and values to the NCI thesaurus and SNOMED CT terminology, what is not clear to me is whether the user is able to select a particular terminology server to use. This is particularly important for local variants of the SNOMED CT terminology that may contain terms not present in the global SNOMED CT terminology.

Furthermore, by providing the functionality to connect to a particular terminology server opens up the possibility for terms from other standardised terminologies and ontologies to be included within the codebooks.

The authors state that “internationally established codebooks, such as CDISC’s CDASH can also be made available”. It would be useful to explicitly state how they propose for this to happen. Will it be posted on a website? Will it be part of the codebook libraries, as is the case with REDCap?
What is also not clear is whether a codebook created for, say, Castor, can be used for another EDC, say OpenClinica. If so, it would be useful to state how the iCRF generator deals with the different implementations (For a start, REDCap implements a slight variation of CDISC ODM, which is not compatible, out-of-the-box, with OpenClinica, which offers a more consistent implementation of CDISC ODM).

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?
Yes

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data science, semantic interoperability and clinical research framework design

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.

Reviewer Report 26 March 2020

https://doi.org/10.5256/f1000research.25280.r61614

Rianne Fijten
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands
Petros Kalendralis
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands

The authors have thoroughly and sufficiently answered any question or concerns we had.
They have added a lot more structure to the article, for example adding a clear use-case scenario in which they describe the steps a user should take (including extra figures for each step).
We are happy to accept this version of the article.

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?
Yes

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

Competing Interests: No competing interests were disclosed.

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.

Version 1

Reviewer Report 14 February 2020

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

?

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 shrinked, a lot of blank space

References

1. Dugas M, Neuhaus P, Meidt A, Doods J, et al.: Portal of medical data models: information infrastructure for medical research and healthcare.*Database (Oxford)*. 2016; **2016**. PubMed Abstract | Publisher Full Text

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.

Author Response 16 Mar 2020

Sander de Ridder, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Thank you reviewing of our paper. Please find below the changes we made to the new version to address your concerns/comments.

Major comments

1. page 2, Methods, Supported Codebooks: Quantitative Information on 6 codebooks is missing: how many item definitions per codebook are available? how many codelist(items)?
We added this information to the “supported codebooks” section.

2. 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
One of the issues is that Castor doesn't support it yet. OpenClinica 3 doesn’t support direct import of ODM CRFs either. Hence, supporting ODM wasn’t a priority for us in the current release, but given your feedback, we added support for CDISC ODM to the roadmap. Accordingly, we updated: "If desirable, additional internationally established formats, such as CDISC ODM, may also be included in the future." in the “Additional EDCs” section to: "Support for CDISC ODM is on the roadmap. If desirable, additional internationally established formats may also be included in the future."

3. 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.
We updated the manuscript to stress that many items have already been mapped.

4. 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 !
As mentioned in an earlier comment, CDISC ODM can be added, but it wasn't a priority for us. However, we added it to the roadmap.

5. 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)
Thank you for the suggestion. We'd have to look into it more closely, but we added it as a possible new feature to the roadmap.

6. 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?
The iCRF Generator is independent of the coding used. Hence, if a codebook there uses UMLS, the iCRF Generator will show UMLS codes. We added the following to the introduction:
Hence, if a codebook is mapped to a medical thesaurus, the iCRF Generator allows the user to use these mappings, preventing the labour-intensive manual mapping.

Minor comments
7. page 2, Introduction, 2nd paragraph - HL7 is not a communication standard, but a standards developing organisation
Health Level Seven International (HL7) is the standards developing organization. However, Health Level Seven (HL7) also refers to the standards. See e.g. their glossary here:
https://www.hl7.org/documentcenter/public/calendarofevents/FirstTime/Glossary%20of%20terms.pdf

Health Level Seven (HL7) is an application protocol for electronic data exchange in health care environments. The HL7 protocol is a collection of standard formats which specify the implementation of interfaces between computer applications from different vendors. This communication protocol allows healthcare institutions to exchange key sets of data amount different application systems. Flexibility is built into the protocol to allow compatibility for specialized data sets that have facility specific needs.

To avoid confusion, we updated the URL in the paper to directly link to the standards (https://www.hl7.org/implement/standards/).

8. page 2, Introduction, 2nd paragraph: the "Meta Data Models website" correct term: Portal of Medical Data Models
We apologise for the error. Fixed.

9. page 2, Introduction, 2nd paragraph: Link to Basic Health Data Set (available contents!) should be updated
Thank for the suggestion. We added the URL.

10. Figure 2 & Figure 4: should be shrinked, a lot of blank space
We removed blank space from the figures.

Competing Interests: No competing interests were disclosed.

Reviewer Report 07 February 2020

https://doi.org/10.5256/f1000research.23777.r59542

?

Rianne Fijten
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands
Petros Kalendralis
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands

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.

Author Response 16 Mar 2020

Sander de Ridder, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Thank you reviewing of our paper. Please find below the changes we made to the new version to address your concerns/comments.

Methods

1. 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.
Indeed, currently the number of codebooks available is limited. We are hoping that by introducing the iCRF Generator to the public, this will stimulate them to publish codebooks in ART-DECOR, which we could then potentially make available in the iCRF Generator. Furthermore, Nictiz is also actively involved; they are currently looking into publishing e.g. established questionnaires.
Given the comments here and later on, we rewrote the “Additional codebooks” paragraph in the discussion to:
At this point, the iCRF Generator gives access to six nationally established codebooks, some of which support multiple languages and multiple versions. To improve the user-base for the iCRF Generator, the number and variety of codebooks available must increase. While nationwide standards, such as the Basic Health Dataset, can be readily made available, some form of governance may have to be put in place to establish other types of curated codebooks. This could stimulate the community to help build high-quality codebooks, mapped to medical thesauri, while preventing too much codebook redundancy and conflicting items. Furthermore, internationally established codebooks, such as CDISC’s CDASH can also be made available.

2. 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.
It makes creating interoperable CRFs easier, not necessarily CRFs. Instead of having to e.g. define Gender and the appropriate codes and manually annotating with the appropriate thesaurus codes, which really takes a lot of time, instead you take the gender from the codebook and reuse that definition. However, you still need to define whether you would like have Gender as a radio button or a dropdown – that work does not change. We checked the manuscript and updated the Use cases section to stress this aspect:
"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."
Rewritten to
"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 manually designing the items and mapping them to a medical thesaurus, which takes a lot of time, the iCRF Generator can be used to select items which have already been mapped from the available codebooks and generate the basis of the case report forms."

3. Figure 1 - In this figure, it is at first not clear where to start. We recommend using numbering to clarify the sequential steps.
New version included in paper in which we numbered the steps

4. 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.
Indeed, it is a valid point that each EDC has its own customisation page, which thereby comes at the cost of maintainability / sustainability. However, this was done to facilitate the data managers.

As an example, Castor's import file is an XML file, which you don't want to edit manually. If we remove all customisation options from de iCRF Generator, this implies all fields will have to be generated with default settings, or the XML will be invalid, e.g. every selection item will be a "required dropdown". This means that a datamanager would import these fields and would then have to manually edit all imported fields in Castor. We believe this would not make for a good user experience.

We added the following to the "EDC-specific item customisation within iCRF Generator" section to clarify our reasoning:
By adding this EDC-specific customisation, the iCRF Generator's code is more difficult to maintain. However, by allowing the data manager to customise essential fields, the iCRF Generator can provide a ready-to-use CRF. This enhances the user experience, making it a worthwhile investment.

5. What does the summary show?
In the text we mention the following about the summary: "The final page of the wizard shows a short summary of the number of selected items." We've added a new Figure showing a screenshot of the summary.

6. 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.
Great idea. We could change the page where you can select the codebooks and make the codebooks themselves e.g. clickable, linking to their ART-DECOR page. We'll add this to the roadmap for a future release.

Use cases
7. We would like to see a clearer example that also shows what each step in figure 1 looks like.
To clarify the relationship between Figure 1 and the example figures (2-6) we updated the text to show how each step in Figure 1 is related to the example figures:

Figure 1 illustrates the iCRF Generator's complete workflow and figure 2-6 show actual examples of the workflow. In a typical use case, the user first selects an EDC from the dropdown (Figure 1, step 1; 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 1, step 2; Figure 3).
<etc>

Discussion
8. The discussion mentions that codebooks might be added when of sufficient quality. How do you define sufficient quality and who decides that?
Our initial aim was to provide a tool which can facilitate the reuse of the available codebooks. How to provide the users with high-quality codebooks is open for discussion and we welcome any input on the matter. We did rewrite this section though - see point 1.

9. 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.
Any codebook can be added to ART-DECOR, as Nictiz welcomes any additions. To facilitate interoperability, we ourselves foremost welcome any (inter)nationally endorsed standard, such as the Basic Health Dataset. We agree that establishing a curated group of codebooks which are properly mapped to medical thesauri would be great, although this does require a proper decision-making structure (governance) as well as the appropriate funding. We rewrote this section - see point 1.

Competing Interests: No competing interests were disclosed.