<RDCIT Data Plan Design>

Purpose of this document: This document details the creation of a RDCIT approved data plan and describes how to integrate clinical coding into a Case Report Form (CRF).

Version: 1.1

Date: 11/03/2016

Authors: Julie von Ziegenweidt, Vera Matser

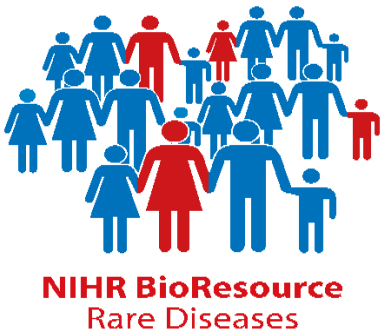
NIHR Rare Diseases Translational Research Collaboration

Department of Haematology, University of Cambridge

National Health Service Blood and Transplant Cambridge

Long Road, Cambridge

CB2 0PT



Version control

Version	Date	Author	Changes
V1.0	18/12/2015	Julie von Ziegenweidt, Vera Matser	First version
V1.1	11/03/2016	Vera Matser	New data plan template (V1.1) with instructions and more examples. Added Ontology Modifiers (section 2.2.7), updated 3.1.4

Check if there is a later version of this document here

Contents

1 Introduction 4

1.1 RDCIT Data Warehouse 4

1.2 Pre-Data Plan Procedures 5

2 Data Plan Design 6

2.1 Variable Description 6

2.2 Clinical Coding 7

2.2.1 RDCIT Data Dictionary 8

2.2.2 RDCIT Controlled Vocabulary (RDCIT-CV) 10

2.2.3 Alternative Ontology Browsers 11

2.2.4 Standard coding 12

2.2.5 Response coding 12

2.2.6 Interpretative coding 13

2.2.7 Ontology Modifiers 14

3 OpenClinica form development 15

3.1 OpenClinica CRF coding rules for data entry forms 16

3.1.1 Standard Coding 16

3.1.2 Response Coding 16

3.1.3 Interpretative Coding 18

3.1.4 Clinically coding multi-select / checkbox without DECODE 21

3.2 OpenClinica CRF coding rules for data import (ODIN) forms 22

3.2.1 Recording of Date of Visit or Date of Lab Test (Importing patient data through ODIN) 22

3.2.2 Standard Coding 23

3.2.3 Response Coding 23

3.2.4 Interpretative Coding 24

4 Preparing data for importing into OpenClinica import forms (CRFs) 27

5 Quality and Code Verification 28

1 Introduction

Developing a Data Plan (DP) is essential to establish an effective method to maximise the potential value of data collected in a research study. Especially if the data is to have significant value as a resource for the wider research community. The Data Plan is what will enable other researchers to access, understand, and utilise study data independently of the original investigators; though data will not be shared without the Principal Investigator's expressed permission.

The first step to developing the Data Plan, is determining what information needs to be collected on the questionnaire or Case Report Form (CRF) and this needs to be based on the specifications of the study protocol. We recommend initially designing the questionnaire on paper and convert it into an electronic (OpenClinica) format once it has been approved. Once that data plan has been approved a final OpenClinica CRF can be produced.

The Data Plan Design Guide should be read in conjunction with the Best Practice Guide as well as the CRF Design Guide. For more background on the importance of clinical coding you can read the RDCIT blog post "Describing Phenotypes using HPO". Instructions on the use of ODIN for importing data into OpenClinica are given in the ODIN User Manual and the Import & Export Guide.

All documentation is available via the RDCIT webpages, under the RDCIT tools and user documentation sections.

1.1 RDCIT Data Warehouse

While reading this guide it will be helpful to be more aware of the RDCIT OpenClinica Data Environment because it underlies and drives the clinical coding procedures laid out in this document. The data environment has three main components (all housed within the RDCIT Secure Data Service); one component (OpenClinica) handles the data entry, the second component is a Reporting Data Warehouse (designated "Raw Study Data") primarily used for mid-study reporting (e.g. recruitment figures) and the third component is the true Data Warehouse (I2B2), which is used for data analysis. Only clinically coded data will be housed in the I2B2 Data Warehouse. Clinical coding dependant on the RDCIT Data Dictionary, as discussed below, is applied before data is moved to the I2B2 Data Warehouse.

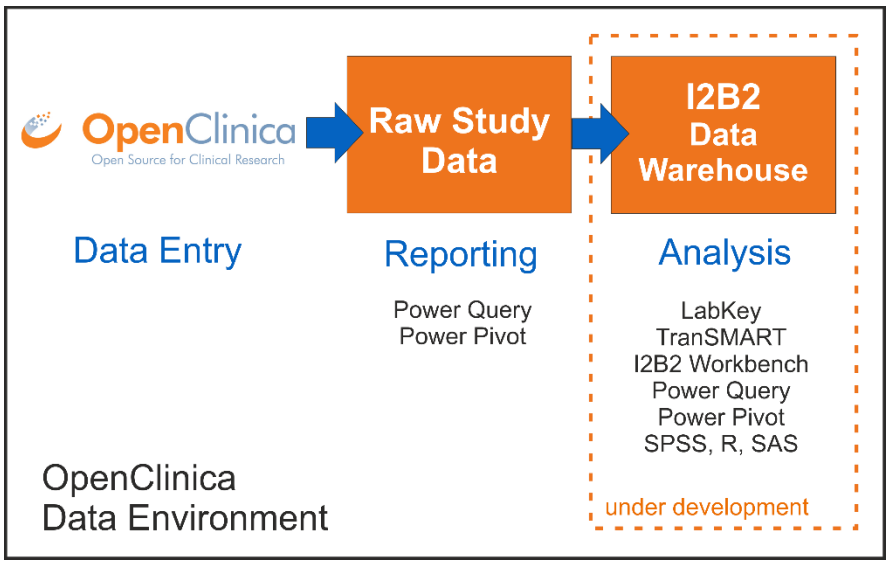


FIGURE 1: SCHEMATIC OVERVIEW OF THE RDCIT OPENCLINICA DATA ENVIRONMENT

1.2 Pre-Data Plan Procedures

Collect all documentation related to the study, such as:

- Identify what patient identifiable information is being recorded
- Data Management Plans (if available)

If the study has already begun or is already completed, also collect the following:

- Paper CRFs
- Documentation on the data items
- Dummy copy of the final data file

Obtain a copy of the final data file which contains headings only. This is to ensure that when the final data file is sent over for verification and incorporation into the OpenClinica database, it is in an import (ODIN) compatible format. If available, provide some anonymous rows of patient data as an indicator of what the data looks like to ensure compliance with the documentation.

2 Data Plan Design

We recommend you use the RDCIT Data Plan Template to design your Data Plan. The most recent version of the data plan template is available on the RDCIT website, under user documentation. Make sure that the Data Plan Design Guide and the Data Plan Template have the same (most recent) version number.

Data Plan Template

The Data Plan template has three sheets; the template sheet where the data plan will be built, an instructions sheet and a sheet with examples to illustrate the data plan design process. Depending on the complexity of your study you may want to copy the template sheet to have a separate sheet per section.

Within the template sheet there is a clear separation between the Variable Description and the Clinical Coding.

2.1 Variable Description

Build the document by starting with the following information (see Table 1):

- **Variable names (Data Item Name)** – indicate the name for each piece of information that is being collected e.g. Gender, Date of birth, Age etc.
- **Variable Description** – provide a brief explanation about the data item being collected. For example, the gender information required might be related to your sex at birth rather than the current sex.
- **Variable Type** – indicate what type of information is being collected; allowable options are string (incl. free text and codes), numeric (incl. integer and real numbers), date, or file.
- **Variable Data Unit types (Unit)** – indicate if there are any specific unit types to be applied for the values entered; e.g. kg for weight, cm for height etc.
- **Variable Response Options** – if there are only to be a set number of response options, please note what the options are; e.g. Yes / No, Male / Female etc.
- **Variable Coded Values** – what format the data will be in when stored in the database i.e. free text or the coded values for the set options for the possible responses, such as ‘M’ for Male / ‘F’ for Female or 1 for Male and 2 for Female.
- **Variable Required** – indicate if the data item is required or not.
- **Variable Rules** – if applicable indicate any business rules required to ensure the quality of the data being collected; e.g. specific date formats, range allowances.

Variable Description							
Data Item Name	Description	Type	Unit	Multi/Single	Response Options	Coded Values	Required
DATE_REG	Date Registered	date	yyyy-mm-dd				Yes
HOSP_NO	Hospital number	numeric					Yes
SEX	Sex	numeric		single	Male	1	Yes
					Female	2	
DATE_DIAG	Date of diagnosis	date	yyyy-mm-dd				Yes
AGE	Age at diagnosis	numeric	years				Yes
HEIGHT	Height at diagnosis	numeric	cm				Yes
WEIGHT	Weight at diagnosis	numeric	kg				Yes
BMI	BMI	numeric					Yes
ASTHMA	Has the patient got Asthma?	numeric		single	Yes	1	Yes
					No	0	

TABLE 1: EXAMPLE OF THE VARIABLE DESCRIPTION SECTION OF THE DATA PLAN TEMPLATE

2.2 Clinical Coding

When developing a Data Plan for a study, it is important to also identify, where possible, what the equivalent clinical coding term is for each of the pieces of data being collected for the participants. This provides a means of standardising and structuring the data for future reporting and analysis.

The clinical coding of the items is vital for enabling the data to be brought into the data warehouse (I2B2) for data mining and cohort identification.

There are multiple ontologies available for clinical coding based on the type of information being collected. Disease specific diagnosis could be managed by using the ICD9 or 10 ontology (International Coding Dictionary established by WHO); HPO (Human Phenotype Ontology) for specific symptom/abnormalities phenotyping; utilising Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) for clinical coding of testing, surgical procedures and therapies. There are also Rare Disease OMIM codes and ORPHANET codes. It is possible to make use of all these ontologies in the same study as long as the type of the ontology is recorded with the code and term. The ontologies that are mainly used for the Rare Disease projects are the Human Phenotype Ontology (HPO) and (SNOMED-CT) dictionaries. Clinical (ontology) coding can be found using ontology browsers available on the internet.

Please Note: Within the Rare Disease projects the Human Phenotype Ontology (HPO) is the preferred ontology for all clinical coding. If however, there is no HPO code is available for the medical term, then utilising the SNOMED-CT ontology is recommended.

Please Note: All clinical ontology (re)coding for data items must be confirmed and signed off by a clinician related to the study who is a specialist in the specific disease area, as well as by the RDCIT Team.

Once the clinician has signed off the ontology documentation, this must be sent to the RDCIT team for study sign off. They will then ensure that all ontology codes are uploaded into the RDCIT Data Dictionary to ensure uniformity across all studies.

The clinical coding discussed in this document has been divided into Standard Coding, Response Coding and Interpretative Coding.

2.2.1 RDCIT Data Dictionary

To improve standardisation across the different projects and for your own convenience, it is highly recommended that you first check the codes which have already been identified by other studies. This provides a quicker method of identifying codes for common data items and for items which are more specific for your study area; subsequently use the online ontology browsers.

The codes already identified are held in the RDCIT Data Dictionary which can be searched by using the online RDCIT Code Browser tool (Figures 2 - 4). Please note that this tool is currently still under development. The RDCIT Code Browser can be accessed through the following address:

URL: http://ocdw.medschl.cam.ac.uk/codebrowser/

The RDCIT Code Browser allows you to search the RDCIT Data Dictionary (RDCIT (all)), HPO or SNOMED-CT. The RDCIT (all) option contain the RDCIT Controlled Vocabulary data items (RDCIT-CV; see below) and the HPO & SNOMED-CT entries already added to the RDCIT Data Dictionary. You can pick your preferred search option in the dropdown menu (Figure 2), though we recommend starting with RDCIT (all).

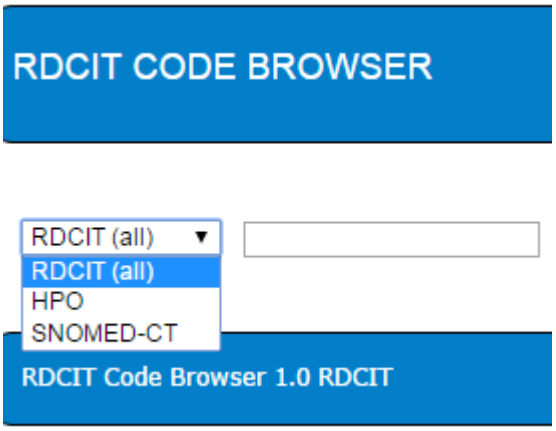


FIGURE 2: RDCIT CODE BROWSER; THE DROPDOWN ALLOWS THE SELECTION OF THE RDCIT (ALL), HPO AND SNOMED-CT

When searching the Code Browser for a term (e.g. Asthma) the results will be displayed in a table (Figure 3).

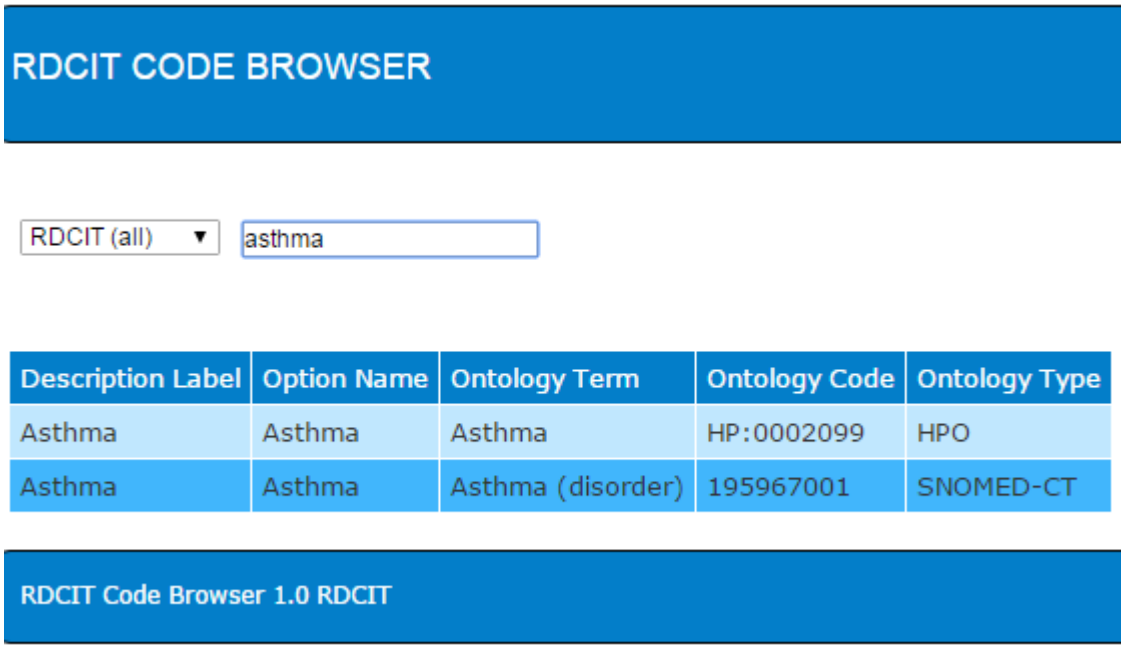


FIGURE 3: EXAMPLE OF RDCIT CODE BROWSER QUERY SCREEN SHOWING THE RESULTS FOR THE SEARCH TERM "ASTHMA"

If you are searching the RDCIT (all) section the *Ontology Type* column will tell you whether the code originates from HPO, SNOMED-CT or is a temporary RDCIT Controlled Vocabulary data items (RDCIT-CV; see below). The *Description Label Ontology Term, Ontology Code* and *Ontology Type* columns are the essential pieces of information to compose the Data Plan. The *Option Name* columns will display additional

information if multiple options are available, if not they will duplicate the *Description Label*. The *Option Name* is best explained with an example such as "Mode of Inheritance"(Figure 4). The *Description Label* is "Mode of Inheritance" because it is the most likely search term when you use the Code Browser. Each Mode of Inheritance option name has their own Ontology Term & Code. Note that Mode of inheritance would be an response coding data item (see section 2.1.5).

RDCIT (all) ▼ mode of inheritance

Description Label	Option Name	Ontology Term	Ontology Code	Ontology Type
Mode of Inheritance	Autosomal dominant contiguous gene syndrome	Autosomal dominant contiguous gene syndrome	HP:0001452	HPO
Mode of Inheritance	Autosomal dominant inheritance	Autosomal dominant inheritance	HP:0000006	HPO
Mode of Inheritance	Autosomal dominant inheritance with maternal imprinting	Autosomal dominant inheritance with maternal imprinting	HP:0012275	HPO
Mode of Inheritance	Autosomal dominant inheritance with paternal imprinting	Autosomal dominant inheritance with paternal imprinting	HP:0012274	HPO
Mode of Inheritance	Autosomal dominant somatic cell mutation	Autosomal dominant somatic cell mutation	HP:0001444	HPO
Mode of Inheritance	Autosomal recessive inheritance	Autosomal recessive inheritance	HP:0000007	HPO

FIGURE 4: SEARCHING FOR A TERM WITH MULTIPLE OPTIONS E.G. MODE OF INHERITANCE (NOTE: NOT ALL OPTIONS ARE SHOWN IN FIGURE)

Please Note: If the ontology term you have selected is not already part of the RDCIT Data Dictionary (but is in HPO, SNOMED-CT or any other recognised ontology) record the Ontology Term and Code in the appropriate column of the Data Plan Template. Please additionally provide a new **Description Label** for the code/term that is descriptive enough but not longer than 20 characters, for other researchers to be able to identify and select for their data items in similar studies. The RDCIT team will automatically add all new codes to the RDCIT Data Dictionary after receiving the completed Data Plan

2.2.2 RDCIT Controlled Vocabulary (RDCIT-CV)

In the event that there is no ontology term that matches the data item you are collecting, or the response you are recording, it is possible to request a new term (& Code) to be created by the RDCIT team.

The RDCIT team has developed an in-house ontology to cover unusual data variables that are not found in standard ontologies but are necessary to code for future analysis; these term are collected in the RDCIT

Controlled Vocabulary (RDCIT-CV). This is a temporary measure until the code becomes available in either the SNOMED-CT or the HPO ontology dictionary.

To request a new code to be introduced, please indicate in the ontology code column, that this is a new code: 'NEW' and in the ontology term column, note down what the new term needs to consist of so that this is available for other researchers to use in their studies. The chosen term must be descriptive enough for the entry to be easily understood and utilized correctly. Please also define a Description Label for the code/term that is descriptive enough but not longer than 20 characters, for other researchers to be able to identify and select for their data items in similar studies.

2.2.3 Alternative Ontology Browsers

If the phenotypic code you are trying to identify for your specific disease area has not been added to the RDCIT Data Dictionary, you can search for the code in the HPO or SNOMED-CT ontologies available in the Code Browser or through the alternative ontology browsers listed below.

2.2.3.1 Human Phenotype Ontology (HPO)

BioPortal - http://bioportal.bioontology.org/ontologies/HP/?p=classes&conceptid=root

MSeqDR HPO Browser - https://mseqdr.org/hpo_browser.php?118

Phenomizer - https://compbio.charite.de/phenomizer/

2.2.3.2 SNOMED-CT

BioPortal - http://bioportal.bioontology.org/ontologies/SNOMEDCT/?p=classes&conceptid=root

NPEx SNOMED-CT Browser - http://www.snomedbrowser.com/

2.2.4 Standard coding

Standard coding items are data items where only a value is collected; e.g. Hospital number, Age or Weight.

To apply standard coding, copy and paste the Clinical Coding information from the RDCIT Code Browser into your data plan template (Table 2):

- Description Label (if already available in the RDCIT Data Dictionary, searchable through RDCIT (all) option in the Code Browser)
- Ontology Term
- Ontology Code
- Ontology Type
- Ontology modifier (if applicable – see section 2.2.7)

Variable Description								Clinical Coding				
Data Item Name	Description	Type	Unit	Multi Single	Response Options	Coded Values	Required	Description Label	Ontology term	Ontology Code	Ontology Type	Ontology Modifier
DATE_REG	Date Registered	date	yyyy-mm-dd				Yes	Recruitment Date	Recruitment Date	RDG00007	RDCIT-CV	
HOSP_NO	Hospital number	numeric					Yes	Hospital Number	Hospital reference number (observable entity)	185975009	SNOMED-CT	
DATE_DIAG	Date of diagnosis	date	yyyy-mm-dd				Yes	Date of diagnosis	Date of diagnosis (observable entity)	432213005	SNOMED-CT	
AGE	Age at diagnosis	numeric	years				Yes	Age at diagnosis	Age at diagnosis (observable entity)	423493009	SNOMED-CT	
HEIGHT	Height at diagnosis	numeric	cm				Yes	Height	Body height measure (observable entity)	50373000	SNOMED-CT	
WEIGHT	Weight at diagnosis	numeric	kg				Yes	Weight	Body weight (observable entity)	27113001	SNOMED-CT	
BMI	BMI	numeric					Yes	BMI	Body mass index (observable entity)	60621009	SNOMED-CT	

TABLE 2: EXAMPLE OF STANDARD CODING FOR DATA ITEMS

2.2.5 Response coding

When a question is asked on the questionnaire (equivalent to OpenClinica CRF) and the result is based on a response value, the individual responses could equate to a diagnosis or symptom of a patient's disease which have a possible matching coded term.

Below is an example of this type of coding using a question on Asthma (Table 3).

A researcher wishes to collect a data item to find out whether the participant has Asthma or not. The only options to this question are going to be a 'Yes' or 'No' response. If the response is 'Yes', then the researcher has decided on the HPO code for Asthma (HP:0002099) to be the mapped ontology code for this question. If the response is 'No' then there is to be no further coding to be done for the Asthma question. The researcher could attach another code in response to the 'No' question if it adds value to the question.

Variable Description								Clinical Coding				
Data Item Name	Description	Type	Unit	Multi/ Single	Response Options	Coded Values	Required	Description Label	Ontology term	Ontology Code	Ontology Type	Ontology Modifier
ASTHMA	Has the patient got Asthma?	numeric		single	Yes	1	Yes	Asthma	Asthma	HP:0002099	HPO	
					No	0						

TABLE 3: EXAMPLE OF CLINICAL CODING FOR RESPONSE VARIABLES USING THE ASTHMA QUESTION

Similar to Standard Coding, add the Clinical Coding information to your Data Plan, adding additional rows for each of the options/response values:

- Description Label (if available in the RDCIT Data Dictionary, searchable through RDCIT (all) option in the Code Browser)
- Ontology Term
- Ontology Code
- Ontology Type
- Ontology Modifier (if applicable – see section 2.2.7)

If a response option does not have an ontology code attached to it (e.g. No/0 for Asthma) the Clinical Coding columns can be left blank. Note that with a data item that has multiple coded options the Description Label will often be the same (see Figure 4 Mode of Inheritance), while the Ontology Term, Code will be different.

2.2.6 Interpretative coding

When recording the result of the lab test or measurement, the researcher could also provide additional interpretative coding to be attached to the same piece of data, thereby enriching the information even further.

In the example shown below, the researcher is planning to see if the antinuclear antibody (ANA) test has been performed on the participant's blood sample and if so, to record the result for the test in the CRF. This lab test measurement can be mapped to the 'Antineutrophil antibody measurement (procedure)' in the SNOMED-CT ontology. In addition to this, if the researcher then records whether the result is Negative or Positive, this can enable additional phenotype coding to be attributed to the Positive response which equates to an HPO term of 'Antineutrophil antibody positivity' which is then attached to the participants data record as well.

Variable Description								Clinical Coding				
Data Item Name	Description	Type	Unit	Multi/ Single	Response Options	Coded Values	Required	Description Label	Ontology term	Ontology Code	Ontology Type	Ontology Modifier
ANA	Antinuclear Antibody	numeric		single	*Negative*	0	Yes					
					Positive	1		ANA positive	Antineutrophil antibody positivity	HP:0003453	HPO	
					Not done	9						
					if not 9 then			ANA	Antineutrophil antibody measurement (procedure)	413563004	SNOMED-CT	

TABLE 4: EXAMPLE OF MULTIPLE CODING FOR A SINGLE DATA ITEM (INTERPRETATIVE CODING)

Similar to Standard and Response Coding, add clinical coding information to your Data Plan, adding additional rows for each of the options/response values.

- Description Label (if available in the RDCIT Data Dictionary, searchable through RDCIT (all) option in the Code Browser)
- Ontology Term
- Ontology Code
- Ontology Type
- Ontology Modifier (if applicable – see section 2.2.7)

If a response option does not have an ontology code attached to it (e.g. Negative/0 or Not done/9 for Antinuclear Antibody) the Clinical Coding columns can be left blank.

2.2.7 Ontology Modifiers

The ontology modifier column is designed to provide an additional way of modifying a code to reflect similar data which is referencing different status or situation for the participant.

For instance, Heart rate can normally only be collected once on a form for a participant.

However, when performing the 6 minute walk test, the heart rate needs to be recorded before the test and after the test. Therefore, instead of creating brand new codes, the same code has been used in 3 different data items in the dictionary:

1st record has an ontology modifier of '@' and description label of 'Heart rate';

2nd record has an ontology modifier of 'PRE:6MWT' and a description label of 'Heart rate - Pre 6mwt' and;

3rd record has an ontology modifier of 'POST:6MWT' and a description label of 'Heart rate - Post 6mwt'.

By changing the description label in the OC forms, the Heart rate can be recorded for the same participant, in the same event and on the same form and will be stored in the data warehouse with the relevant modifier to indicate the differences.

Variable Description								Clinical Coding				
Data Item Name	Description	Type	Unit	Multi/ Single	Response Options	Coded Values	Required	Description Label	Ontology term	Ontology Code	Ontology Type	Ontology Modifier
Heart rate	Heart rate of participant	numeric					Yes	Heart rate	Heart rate (Observable entity)	364075005	SNOMED-CT	@
Heart rate Pre 6mwt	Heart rate pre 6min walk test	numeric					Yes	Heart rate - Pre 6mwt	Heart rate (Observable entity)	364075006	SNOMED-CT	PRE:6MWT
Heart rate Post 6mwt	Heart rate post 6min walk test	numeric					Yes	Heart rate - Post 6mwt	Heart rate (Observable entity)	364075007	SNOMED-CT	POST:6MWT

TABLE 5: USING AN ONTOLOGY MODIFIER ILLUSTRATED WITH THE EXAMPLE OF COLLECTION HEART RATE DURING THE 6 MINUTE WALK TEST.

3 OpenClinica form development

After the Data Plan has been developed and signed off by the clinician in charge as well as the RDCIT team the OpenClinica Case Report Form (CRF) can be developed. Almost all information necessary to develop a CRF is already contained within the Data Plan. We make a distinction in the development of a CRF that is meant for live data entry and those meant for data import using the OpenClinica Data Importer web applicatioN (ODIN). The sections below detail how the clinical coding that was determined in the Data Plan can be implemented in an OpenClinica CRF.

Please Note: For details on OpenClinica CRF design refer to the RDCIT CRF Design Guide.
http://rdcit.org/openclinica/oc-training/user-documentation/

3.1 OpenClinica CRF coding rules for data entry forms

Please Note: Match the **Description Label** from the Data plan spreadsheet to the **DESCRIPTION LABEL** column in the CRF to automatically map this field to the actual code item.

3.1.1 Standard Coding

For Standard Coding data items, it is only necessary to create a data item in the CRF for recording the response of the question being asked or value to be captured. The example given below is for the Data Item *Gender.*

ITEM_NAME	GENDER
DESCRIPTION_LABEL	Gender
LEFT_ITEM_TEXT	Sex at birth
RESPONSE_TYPE	radio
RESPONSE_LABEL	M_F
RESPONSE_OPTIONS_TEXT	Male, Female
RESPONSE_VALUES_OR_CALCULATIONS	1,2
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 6: REPRESENTATION OF THE CRF STRUCTURE FOR STANDARD CODING DATA ITEMS

3.1.2 Response Coding

To add Response Coding to a CRF a minimum of two data items will be required. We will illustrate based on the example data item *Asthma*.

- **Data Item 1:** Create the data item in the CRF for recording the response of the question being asked or value to be captured (similar to Standard Coding data item)

The first data items in the CRF will look as follows (e.g. Asthma):

ITEM_NAME	ASTHMA_Q1

DESCRIPTION_LABEL	Asthma
LEFT_ITEM_TEXT	Has the patient got Asthma?
RESPONSE_TYPE	radio
RESPONSE_LABEL	Y_N
RESPONSE_OPTIONS_TEXT	Yes,No
RESPONSE_VALUES_OR_CALCULATIONS	1,0
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 7: REPRESENTATION OF THE CRF STRUCTURE FOR A RESPONSE CODING DATA ITEM TO CAPTURE THE ACTUAL VALUE (DATA ITEM 1)

- **Data Item 2:** Create a second data item in the CRF for capturing the related ontology interpretation code for that data item. The best method to capture the code within a form item which is not entered directly by the user, is by the use of the DECODE function.

The second CRF data item will need to have the following values:

RESPONSE TYPE = Calculation
RESPONSE LABEL = Calc1 (this is any label that you choose to use. Must be unique per calculation)
RESPONSE OPTIONS TEXT = Calc1 (this must match the response label)
RESPONSE_VALUES_OR_CALCULATIONS = func: DECODE()
DATA_TYPE = ST (string)
ITEM_DISPLAY_STATUS = HIDE (Coded data items are normally hidden as not needed to be shown on the form.)

In the case of our *Asthma* example the CRF structure for the second data item will look as follows:

ITEM_NAME	ASTHMA_HPO
DESCRIPTION_LABEL	Asthma
LEFT_ITEM_TEXT	Asthma_HPO
RESPONSE_TYPE	calculation
RESPONSE_LABEL	Calc1
RESPONSE_OPTIONS_TEXT	Calc1
RESPONSE_VALUES_OR_CALCULATIONS	func: DECODE (Asthma_Q1,1, HP:0002099)

DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 8: REPRESENTATION OF THE CRF STRUCTURE FOR A RESPONSE CODING DATA ITEM FOR CAPTURING THE INTERPRETATION OF THE RESPONSE (DATA ITEM 2)

OpenClinica will ask the question 'Has the participant got Asthma?' on the form which provides a radio button for selecting either 'Yes' or 'No'. Whichever option is selected, will have its specified value stored in the data item for Asthma_Q1. If 'Yes' is selected, then the HPO phenotype code HP:0002099 is stored in the data item for Asthma_HPO.

3.1.3 Interpretative Coding

To add Interpretative Coding to a CRF a minimum of three data items will be required. We will illustrate based on the example data item *Antinuclear Antibody (ANA).*

- **Data Item 1:** Create the data item in the CRF for recording the response of the question being asked or value to be captured (similar to Standard Coding data item).

ITEM_NAME	ANA_Q2
DESCRIPTION_LABEL	ANA
LEFT_ITEM_TEXT	ANA Test?
RESPONSE_TYPE	radio
RESPONSE_LABEL	N_P_ND
RESPONSE_OPTIONS_TEXT	Negative,Positive,Not Done
RESPONSE_VALUES_OR_CALCULATIONS	0,1,9
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 9: REPRESENTATION OF THE CRF STRUCTURE FOR AN INTERPRETATIVE CODING DATA ITEM TO CAPTURE THE RESPONSE TO THE QUESTION (DATA ITEM 1)

- **Data Item 2:** Create a second data item in the CRF for capturing the related ontology interpretation code for that lab test data item (similar to Response Coding data item). The best method to capture the code within a form item which is not entered directly by the user, is by the use of the DECODE function. The decode function works in the logical evaluation of an 'If, Else statement'.
  (If 1, do something, else if 2, do something, else if 3, do something, otherwise use this default value)

Set the second CRF data item to the following values:

RESPONSE TYPE = Calculation
RESPONSE LABEL = Calc1 (this is any label that you choose to use. Must be unique per calculation)
RESPONSE OPTIONS TEXT = Calc1 (this must match the response label)
RESPONSE_VALUES_OR_CALCULATIONS = func: DECODE()
DATA_TYPE = ST (string)
ITEM_DISPLAY_STATUS = HIDE (Coded data items are normally hidden as not needed to be shown on the form.)

In the case of our *Antinuclear Antibody (ANA)* example the CRF structure for the second data item will look as follows:

ITEM_NAME	ANA_CODE
DESCRIPTION_LABEL	ANA
LEFT_ITEM_TEXT	ANA_CODE
RESPONSE_TYPE	Calculation
RESPONSE_LABEL	Calc1
RESPONSE_OPTIONS_TEXT	Calc1
RESPONSE_VALUES_OR_CALCULATIONS	func: DECODE (ANA_Q2,1, 413563004,0, 413563004)
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 10: REPRESENTATION OF THE CRF STRUCTURE FOR AN INTERPRETATIVE CODING DATA ITEM FOR CAPTURING THE INTERPRETATION OF THE RESPONSE (DATA ITEM 2)

The DECODE function is going to store the SNOMED-CT **413563004** code in this data item if either Negative option (0) or the Positive (1) was selected. If the Not Done option (9) was selected, then the data item will remain blank.

- **Data Item 3:** Create a third data item in the CRF for capturing the additional ontology interpretation code based on the response. The best method to capture the code within a form item which is not entered directly by the user, is by the use of the DECODE function. The decode function works in the logical evaluation of an If, Else statement.
  (If 1, do something, else if 2, do something, else if 3, do something, otherwise use this default value)

Set the third CRF item to the following values:

RESPONSE TYPE = Calculation
RESPONSE LABEL = Calc2 (this is any label that you choose to use. Must be unique per calculation)
RESPONSE OPTIONS TEXT = Calc2 (this must match the response label)
RESPONSE_VALUES_OR_CALCULATIONS = func: DECODE()
DATA_TYPE = ST (string)
ITEM_DISPLAY_STATUS = HIDE (Coded data items are normally hidden as not needed to be shown on the form.)

In the case of our *Antinuclear Antibody (ANA)* example the CRF structure of the third data items would look as follows:

ITEM_NAME	ANA_HPO
DESCRIPTION_LABEL	ANA Positive
LEFT_ITEM_TEXT	ANA Positive HPO
RESPONSE_TYPE	Calculation
RESPONSE_LABEL	calc2
RESPONSE_OPTIONS_TEXT	calc2
RESPONSE_VALUES_OR_CALCULATIONS	func: DECODE (ANA_Q2,1, HP:0003453)
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 11: REPRESENTATION OF THE CRF STRUCTURE FOR AN INTERPRETATIVE CODING DATA ITEMS FOR CAPTURING THE HPO INTERPRETATION OF THE PHENOTYPE (DATA ITEM 3)

The Decode function is going to store the HPO phenotype code **HP:0003453** code in this data item if the Positive option (1) was selected. If the Negative option (0) or the Not Done option (9) was selected, then the data item will remain blank.

OpenClinica will ask the question 'ANA Results?'' with a radio button for either selecting 'Negative' or 'Positive' or 'Not Done'. Whichever option is selected, will have its specified value stored in the data item for ANA_Q2. If 'Yes' is selected, then the HPO phenotype code HP:0002099 is stored in the data item for ANA_HPO.

3.1.4 Clinically coding multi-select / checkbox without DECODE

Multi-select and checkboxes (especially with many response options) are difficult and time consuming to implement using the DECODE function. In these cases it is possible to clinically code data added to the CRF without using the DECODE function. If it is possible to add the Ontology Term to the RESPONSE_OPTIONS_TEXT column of the CRF and the Ontology Code to the RESPONSE_VALUES_OR_CALCULATIONS, you can code items without using DECODE even if they are not standard coding items.

> **Please Note:** Note that this should not be implemented for single select and avoided if possible because it complicates the downstream analysis in the data warehouse.

We will illustrate this based on the *Mode of Inheritance* (see Figure 4) example. Note that for space reasons only the first 2 options will be added to the table below, however, there is no limit on how many options you can add in this fashion.

ITEM_NAME	MODE_INHER
DESCRIPTION_LABEL	Mode of Inheritance
LEFT_ITEM_TEXT	Mode of Inheritance
RESPONSE_TYPE	single-select
RESPONSE_LABEL	ssl1
RESPONSE_OPTIONS_TEXT	Autosomal dominant contiguous gene syndrome, Autosomal dominant inheritance
RESPONSE_VALUES_OR_CALCULATIONS	HP:0001452, HP:0000006
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	

TABLE 12: CLINICALLY CODING MULTI-SELECT / CHECKBOX WITHOUT USING DECODE

3.2 OpenClinica CRF coding rules for data import (ODIN) forms

Please Note: Match the **Description Label** from the data plan spreadsheet to the **DESCRIPTION LABEL** column in the CRF to automatically map this field to the actual code item.

Please Note: The DECODE function does not trigger when data is imported into the OpenClinica form! Therefore, the code needs to be manually entered within the actual rows of data being imported so the CRF only needs to have a data item available to hold the value.

3.2.1 Recording of Date of Visit or Date of Lab Test (Importing patient data through ODIN)

Please Note: When coding for data that has already been collected, and therefore capturing the actual date of visit or date of lab results in the CRF, it is recommended that the RDCIV-CV 'RDG00019', term of 'Date of Clinical Recording' is coded against the visit date data item.

Coding against 'RDG00019 - Date of Clinical Recording' will ensure that when the data items are created in the data warehouse with their specified codes, the date of entry for these items will be indicated as the original date of visit or original date of the lab test, instead of the default (when importing data), which is the date of the scheduled event.

For example: when importing the data through ODIN, the date of the information of that event for the CRF, will be the scheduled event date in the study i.e. today's date.

However, the actual date of the visit, could have been a year ago and could be different for each visit and for each patient.

Therefore, by coding the date of the visit in the patients record against 'RDG00019', the system will know that when transferring the data into the data warehouse, the actual date of the information being collected, is not the 'event date' but rather the date held within the value of the RDG00019 code.

3.2.2 Standard Coding

For Standard Coding it is only necessary to create a data item in the CRF for recording the response of the question being asked or value to be captured. The example given below is for the Data Item *Gender*.

ITEM_NAME	GENDER
DESCRIPTION_LABEL	Gender
LEFT_ITEM_TEXT	Sex at birth
RESPONSE_TYPE	radio
RESPONSE_LABEL	M_F
RESPONSE_OPTIONS_TEXT	Male, Female
RESPONSE_VALUES_OR_CALCULATIONS	1,2
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 13: REPRESENTATION OF THE IMPORT CRF STRUCTURE FOR STANDARD CODING DATA ITEMS

3.2.3 Response Coding

To add Response Coding to a data import CRF a minimum of two data items will be required. We will illustrate based on the example data item *Asthma*.

- **Data Item 1:** Create the data item in the import CRF for recording the response of the question being asked or value to be captured.

The import CRF structure of the first data items in the CRF would look as follows:

ITEM_NAME	ASTHMA_Q1
DESCRIPTION_LABEL	Asthma
LEFT_ITEM_TEXT	Has the patient got Asthma?
RESPONSE_TYPE	radio

RESPONSE_LABEL	Y_N
RESPONSE_OPTIONS_TEXT	Yes,No
RESPONSE_VALUES_OR_CALCULATIONS	1,0
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 14: REPRESENTATION OF THE IMPORT CRF STRUCTURE FOR A RESPONSE CODING DATA ITEM TO CAPTURE THE ACTUAL VALUE (DATA ITEM 1)

- **Data Item 2:** Create a second data item in the import CRF for capturing the related ontology interpretation code for that data item.

The second CRF data item will need to have the following values:

ITEM_NAME	ASTHMA_HPO
DESCRIPTION_LABEL	Asthma
LEFT_ITEM_TEXT	Asthma_HPO
RESPONSE_TYPE	Text
RESPONSE_LABEL	Leave blank
RESPONSE_OPTIONS_TEXT	Leave blank
RESPONSE_VALUES_OR_CALCULATIONS	Leave blank
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 15: REPRESENTATION OF THE IMPORT CRF STRUCTURE FOR A RESPONSE CODING DATA ITEM FOR CAPTURING THE INTERPRETATION OF THE RESPONSE (DATA ITEM 2)

3.2.4 Interpretative Coding

To add Interpretative Coding to an import CRF a minimum of three data items will be required. We will illustrate based on the example data item *Antinuclear Antibody (ANA)*.

- **Data Item 1:** Create the data item in the import CRF for recording the response of the question being asked or value to be captured.

The first data item in the import CRF looks as follows:

ITEM_NAME	ANA_Q2
DESCRIPTION_LABEL	ANA
LEFT_ITEM_TEXT	ANA Test?
RESPONSE_TYPE	radio
RESPONSE_LABEL	N_P_ND
RESPONSE_OPTIONS_TEXT	Negative,Positive,Not Done
RESPONSE_VALUES_OR_CALCULATIONS	0,1,9
DATA_TYPE	INT
ITEM_DISPLAY_STATUS	SHOW

TABLE 16: REPRESENTATION OF THE IMPORT CRF STRUCTURE FOR AN INTERPRETATIVE CODING DATA ITEM TO CAPTURE THE RESPONSE TO THE QUESTION (DATA ITEM 1)

- **Data Item 2:** Create a second data item in the import CRF for capturing the related ontology interpretation code for that lab test data item.

The second data item in the import CRF looks as follows:

ITEM_NAME	ANA_CODE
DESCRIPTION_LABEL	ANA
LEFT_ITEM_TEXT	ANA_CODE
RESPONSE_TYPE	Text
RESPONSE_LABEL	Leave blank
RESPONSE_OPTIONS_TEXT	Leave blank
RESPONSE_VALUES_OR_CALCULATIONS	Leave blank
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 17: REPRESENTATION OF THE IMPORT CRF STRUCTURE FOR AN INTERPRETATIVE CODING DATA ITEM FOR CAPTURING THE INTERPRETATION OF THE RESPONSE (DATA ITEM 2)

- **Data Item 3:** Create a third data item in the import CRF for capturing the additional ontology interpretation code based on the response.

The third data items in the import CRF looks as follows:

ITEM_NAME	ANA_HPO
DESCRIPTION_LABEL	ANA Positive
LEFT_ITEM_TEXT	ANA Positive HPO
RESPONSE_TYPE	Text
RESPONSE_LABEL	Leave blank
RESPONSE_OPTIONS_TEXT	Leave blank
RESPONSE_VALUES_OR_CALCULATIONS	Leave blank
DATA_TYPE	ST
ITEM_DISPLAY_STATUS	HIDE

TABLE 18: DATA ITEM 3 FOR CAPTURING THE HPO INTERPRETATION OF THE PHENOTYPE

4 Preparing data for importing into OpenClinica import forms (CRFs)

Please Note: Refer to the ODIN user manual and the RDCIT Import & Export Guide for full guidance on how to import data into OpenClinica. The documentation is available on the RDCIT webpages:
ODIN User Manual - http://rdcit.org/openclinica/rdcit-tools/
RDCIT Import & Export Guide - http://rdcit.org/openclinica/oc-training/user-documentation/

The RDCIT has developed a browser-based tool that allows users to bulk-import data into OpenClinica. The raw data has to be prepared to be able to import.

Step 1: Prepare your raw dataset for import; the format of the raw dataset has to adhere to the ODIN data template, additionally we advise users to align their raw data structure to the CRF structure (or vice versa). Having a similar structure and heading will make the import process (especially the mapping step) significantly easier.
Step 2: As the DECODE function in the CRF does not trigger when data is imported into the form, the relevant clinical codes must now be added into the raw dataset prior to importing.

Using the *Asthma* example in Table 3, the clinical codes need to be added to the raw dataset in order to be part of the imported dataset. In this example the *Asthma* column is coded 1 for "Yes" and 0 for "No".

PATIENT No.	ASTHMA	*ASTHMA_HPO*
1	1	*HP:0002099*
2	0	
3	0	
4	1	*HP:0002099*

TABLE 19: RAW DATASET CODING ASTHMA DATA ITEMS AFTER ADDING CLINICAL CODING INFORMATION TO THE RAW DATA SET (NOTE: THIS EXAMPLE IS NOT IN THE ODIN TEMPLATE).

In the example coding *Antinuclear Antibody* data items (Table 4), the ANA data items in the raw dataset could look like the following:

PATIENT No.	ANA_Result	*ANA_Code*	*ANA_HPO*
1	1	*413563004*	*HP:0003453*
2	0	*413563004*	

3	0	*413563004*	
4	9		

TABLE 20: EXAMPLE OF A RAW DATASET CODING ANTINUCLEAR ANTIBODY DATA ITEMS. THE ANA_RESULT COLUMN IS CODED 1 FOR POSITIVE, 0 FOR NEGATIVE AND 9 FOR NOT DONE (NOTE: THIS EXAMPLE IS NOT IN THE ODIN TEMPLATE)

When the test is not performed, then do not record the SNOMED-CT code, as this should only be used to indicate that the test was actually performed for the participant. If the test result is inconclusive, e.g. 'Unknown', then it's still important to record that the test has been done but result is 'Unknown'.

5 Quality and Code Verification

Please Note: All clinical ontology (re)coding for data items must be confirmed and signed off by a clinician related to the study who is a specialist in the specific disease area.

Once the data has been imported into the OpenClinica study, the data needs to be exported for checking & verification.

The following checks need to be completed:

a) The number of patients in the original dataset must match the number of subjects which have been created in the study.
b) Check that the demographics match the details of the subjects in the original dataset.
c) Verify that any code mappings have been interpreted correctly.
d) Ensure that the ontology codes that are produced from the calculations are the correct codes to represent the data being collected.