

Common Data Elements Generator (CDEGEN): A Basic Instruction Guide

By Julian Varghese, MD, MSc

Email: Julian.Varghese@uni-muenster.de

Background

This guide summarizes a user scenario, where a user selects multiple medical data models for Myeloid Leukemia on the Medical Data Models Portal (MDM-Portal): <https://medical-data-models.org/>. From there, selected forms can be loaded into the Common Data Elements Generator (CDEGen) with full support of semantic analyses and common data element generation. This is the recommended way of selecting and loading data models into CDEGen and this guide will focus on this approach.

However, on <https://cdegenerator.uni-muenster.de/> the user may select his own Medical forms as CDISC ODM files or Excel templates with UMLS coding. Semantic analyses will only work if the files are UMLS-coded. If you wish to UMLS-code your own ODM files, we can assist you to do so or you enter the UMLS-codes via ODM-Edit Webtool on the MDM-Portal (After registration, go to Actions -> ODM-Edit Tool).

Step 1) Get started: Register and find appropriate Medical Data Models to compare

Go to <https://medical-data-models.org/> and register for a free account. If you are already registered, login with your account (see Figure 1)

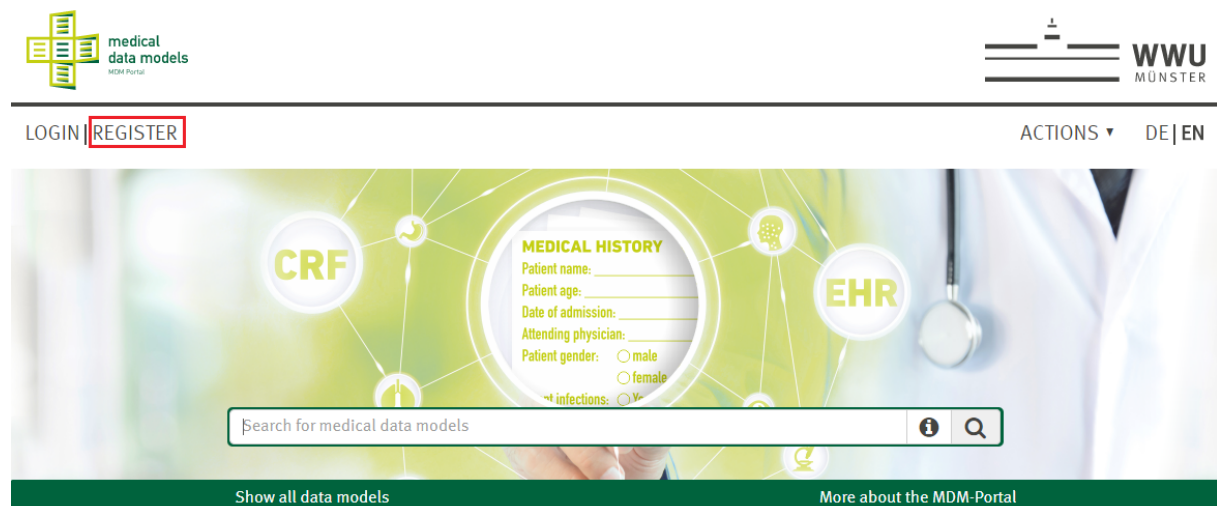


Figure 1. View for unregistered or logged-out user. Login or Register to the MDM-portal and CDEGen.

Then, type into the search field: myeloid leukemia. A number of different MeSH terms or other possible search matches will be suggested (see Figure 2). However, no appropriate suggestion is listed, therefore we will keep our term “myeloid leukemia” and will click on the search symbol (red rectangle in Figure 2).

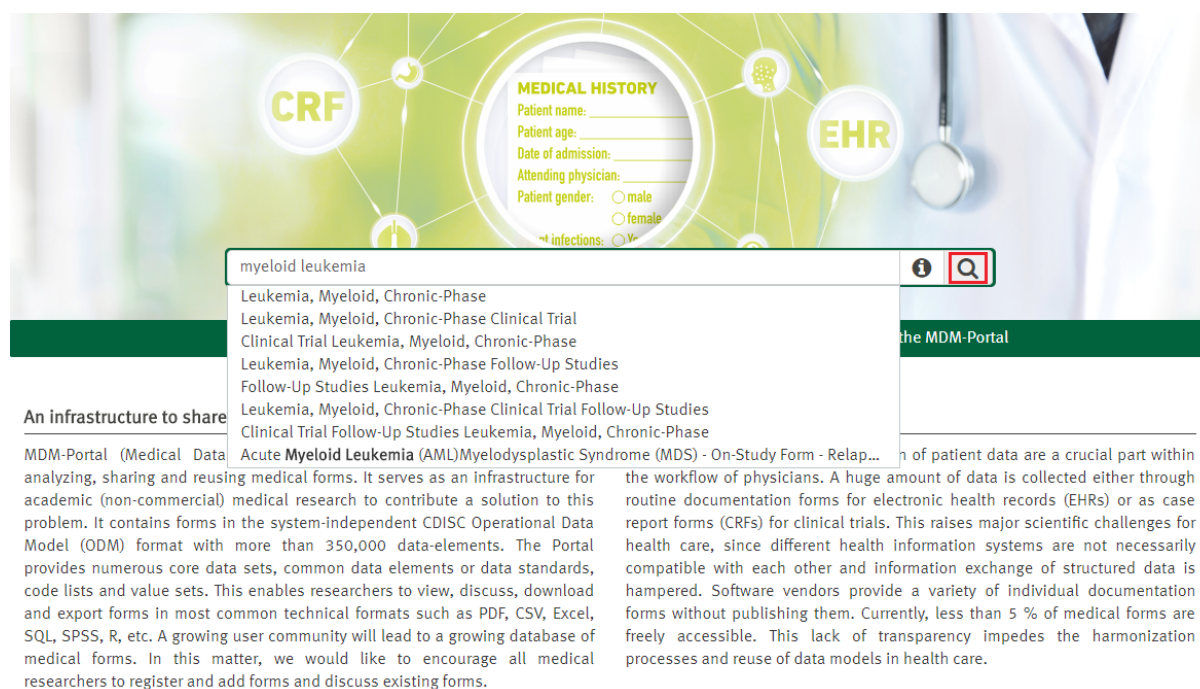


Figure 2. Autosuggestion of controlled terms (don't need to be selected), press the magnifier symbol to start the search.

A list of possible search results will be listed (Figure 3). The search can be further refined by selecting keywords or specific items in the table of contents on the left navigation section. You may also click on the title of a search result to view its detailed item groups and items. To get back to the search results you may hit the browser return button. Following the steps 1.-6. in Figure 3 will result in adding the four top-listed data models into the comparison list:

Each search result (listed in the center of the website) has plus-sign  right next to it. Click it (Figure 3. Steps 1.-4.) to add the models into the comparison list (Figure 3. Step5.) and click "Analyze in CDEGenerator" to start analysis (Figure 3. Step 6.).

myeloid leukemia

ACTIONS ▾

DE|EN

Sort (Relevance) ▾

1352 Search results.

Eligibility Leukemia, Myeloid NCT00512083

★★★★★ - 01-10-2018 - 1 form, 2 itemgroups, 9 items, 1 language

Itemgroups: Inclusion Criteria, Exclusion Criteria

Phase II Study of AS1411 Combined With Cytarabine to Treat Acute Myeloid Leukemia; ODM derived from...

Clinical Trial · Eligibility Determination · Leukemia, Myeloid, Acute

Eligibility Leukemia, Myeloid NCT00611247

★★★★★ - 03-07-2017 - 1 form, 2 itemgroups, 20 items, 1 language

Itemgroups: Inclusion Criteria, Exclusion Criteria

Phase 2 Study of Temozolomide to Treat Poor Risk / Refractory Acute Myeloid Leukemia; ODM derived from...

Eligibility Determination · Hematology · Leukemia, Myeloid, Acute

Eligibility Leukemia, Myeloid NCT00512083

☆☆☆☆☆ - 01-10-2018 - 1 form, 2 itemgroups, 10 items, 1 language

Itemgroups: Inclusion Criteria, Exclusion Criteria

Phase II Study of AS1411 Combined With Cytarabine to Treat Acute Myeloid Leukemia; Eligibility ODM d...

Eligibility Determination · Clinical Trial · Leukemia, Myeloid, Acute

26672_Common Data Elements Myeloid Leukemia (MDS AML CML) - Global Disease Course

★★★★☆ - 01-16-2018 - 13 forms, 3 itemgroups, 11 items, 1 language

Itemgroups: Global Disease Course, Red Cell Transfusion, Platelet Cell Transfusion

Common Data Elements Myeloid Leukemia (MDS, AML, CML) for use in clinical routine and clinical trial ...

Common Data Elements (CDE) · Draft Standard · Leukemia

26672_Common Data Elements Myeloid Leukemia (MDS AML CML) - Physical Examination/Follow up

★★★★☆ - 01-16-2018 - 13 forms, 3 itemgroups, 14 items, 1 language

Itemgroups: Physical Examination/Follow up, Vitals, CTCAE

Common Data Elements Myeloid Leukemia (MDS, AML, CML) for use in clinical routine and clinical trial ...

Common Data Elements (CDE) · Draft Standard · Leukemia

26672_Common Data Elements Myeloid Leukemia (MDS AML CML) - Administrative/Demographic Details

★★★★☆ - 01-16-2018 - 13 forms, 1 itemgroup, 11 items, 1 language

Itemgroup: Administrative/Demographic Details

Common Data Elements Myeloid Leukemia (MDS, AML, CML) for use in clinical routine and clinical trial ...

Common Data Elements (CDE) · Draft Standard · Leukemia

Filter search results

Keywords

Clinical Trial (1162)

Eligibility Determination (421)

Leukemia (347)

AML (247)

Hypereosinophilic Syndr... (199)

Leukemia, Myeloid, Acute (164)

HIV (126)

Hematology (115)

Show more Keywords

Table of contents

1. Clinical Trial

2. Routine Documentation

3. Registry/Cohort Study

4. Quality Assurance

5. Data Standard

6. Patient-Reported Outcome

7. Medical Specialty

Select data models

Selected data models

Eligibility Leukemia, Myeloid NCT...

Eligibility Leukemia, Myeloid NCT...

Eligibility Leukemia, Myeloid NCT...

26672_Common Data Elements M...

Download selection

Analyze in CDEGenerator

Figure 3. Steps to add data models (1-4.) to the selection list (5.). Click on 6. to import all of the selected models to CDEGen (Figure 4)

The user will be directed to webpage of CDEGen with the selected input files (Figure 4). CDEGen is configured with recommended standard settings, which should only be changed if the user is well-experienced in the use of CDEGen. Click on the button *Start Analysis*.

Step 2) Start Semantic Analyses

Common Data Elements Generator

The screenshot shows the 'Common Data Elements Generator' interface with five steps:

- 1. Select input files:** Includes 4 ODM-Files vom MDM-Portal.
- 2. Select input options:** ☒ Extract from ItemDef, ☐ Extract from CodeListItem, ☐ Extract from ItemGroupDef.
- 3. Select output options:** ☒ Pre-coordinated concepts: show partial overlaps, ☐ Post-coordinated concepts: show partial overlaps, ☒ Get UMLS notations.
- 4. Select output format:** ☒ Output HTML, ☐ Output Excel.
- 5. Generate Output:** A blue button labeled 'Start Analysis' is highlighted with a red rectangle.

Figure 4. Starting page of CDEGen. New user are recommended to remain the default setting and to proceed the analysis by clicking on "Start Analysis".

CDEGen Analysis Options (Figure 4):

1. Select input files: The files were already selected by the MDM-portal, therefore no further user input is required.

2. Select input options: By default, only concepts coded in the ItemDef elements of the CDISC ODM files will be regarded in the analyses. However, Concepts from CodeListItems and whole ItemGroups could also be compared to each other if they were UMLS coded in the input sources. On the MDM-portal, UMLS coding is mostly supported for ItemDefs and ItemGroupDefs

3. Select output options

Pre-coordinated concepts: show partial overlaps (default active): Partial overlaps between pre-coordinated and post-coordinated concepts will be shown. Example: Item1 has following coding: C1 (pre-coordinated), Item 2 has following coding: C1 C3 (post-coordinated). Since C1 is common in both items (partial overlap), this will be shown in the analyses. It has to be noted, that partial overlaps are only based on same UMLS codes and not similarity-based on ontological hierarchies or relations.

Post-coordinated concepts: show partial overlaps: Partial overlaps of two post-coordinated concepts will be shown. In default, this option is de-activated, since activation would show a large number of overlapping concepts though only one irrelevant UMLS code is in common.

Get UMLS notations (default active): Show official UMLS-based descriptions and names of concepts. This option establishes a link to the UMLS-metathesaurus database and might impair run time performance if large input files (eg. Containing more than 1000 items) were selected.

4. Select output format

Default option is HTML this guide will rely on. To process results with Excel or further statistic programs an Excel-based output of raw results-data is also possible.

Step 3) View Analyses Results

After analysis is started, analysis results are structured from top to down into three essential sections (A-C) and a last section illustrating the frequency sorted list of concepts.

A: Analyses modes (Fig 5.A)

- Mode: *All concept occurrences*. The default mode, which will be initially shown. Semantic overlaps will be only based on same UMLS coding.
More Details:
All following analyses features in B - D will apply to all UMLS coded concepts, of all documentation hierarchies: Concepts coded in Itemdefs , ItemGroups and CodeListItems (if selected in Step 2, under input options) .
It has to be noted, that a concept can also contain multiple interrelating post-coordinated codes. In this case each concept will be shown as a set of the inter-related UMLS codes.
- Mode: *Matching data elements*: B-C Analyses feature will apply only to ItemDefs. This is a special analysis option dedicated to Itemdefs, since an Itemdef shares a standardized set of attributes to compare with (eg, name, question label, datatype, Measurement Unit, UMLS code). Semantic Matches are shown based on same UMLS coding (as mode *All concept occurrences*) plus same datatypes.
More details:
If an Itemdef is linked to Codelist, a matching Itemdef must also be linked to a semantically equivalent Codelist, which shares the same set of UMLS coded concepts for each permissible value in Codelist (=CodeListItem).
- Mode *identical data elements*: As in mode *Matching data elements*, additionally Itemdefs need to share the same measurement unit (if available), same item question (if same language).

B: Cumulative Frequency Analyses (Fig 5.B), see Figure for description.

C: Overlap Matrix (Fig 5.C), see Figure for description

[Back to Input](#)

A: Analyses modes

CDEGenerator

Download

All concept occurrences

Matching data elements

Identical data elements

All concept occurrences

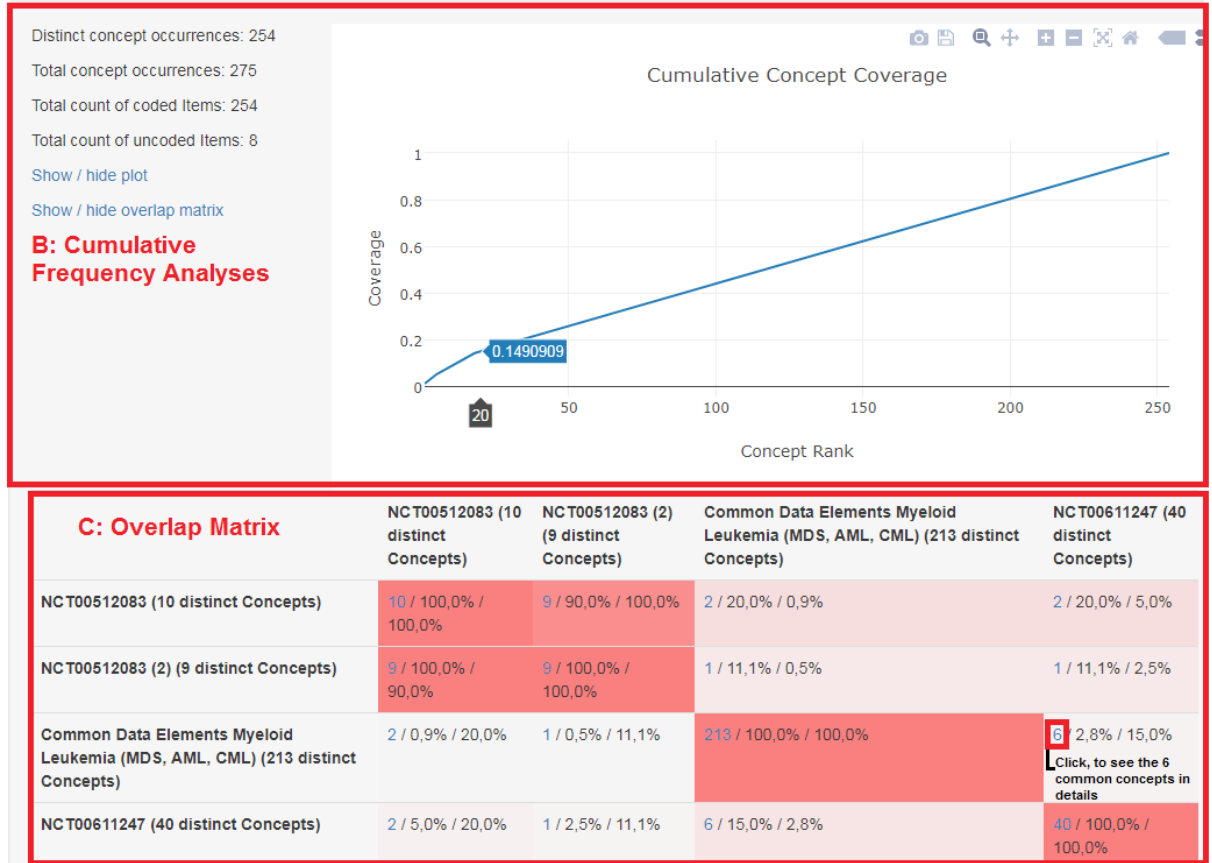


Fig 5: A: Select the Analyses mode, all concept occurrences mode is the default one.

B: Interactive Cumulative Frequency Analysis (blue solid graph). By hovering along the X-axis, the user can choose the set size of the most frequent concepts and immediately sees coverage of all concept occurrences. As shown here, the 20 most common concepts cover 14,9% of all 254 concept occurrences. Initial high slope of the graph indicates existence of highly repetitive concepts within the sources.

C: Heatmap-based similarity matrix of selected sources. Each cell contains the number of common concepts of two sources. Two additional numbers provide percentages that represent the relative overlap in source 1 and source 2. For instance, the bottom right cell with the label: *click to see the 6 common concepts in details* illustrates concept overlaps between the common data elements source and eligibility criteria of study NCT00611246. There are six common concepts, which can be reviewed in detail in the frequency-sorted list of concepts upon clicking (see Fig 6). Since the first source 213 concepts and the second 40, the relative overlap is higher in the second source (2.8% vs 15%).

Frequency-sorted list of concepts (Fig 6)

- As part of the starting analysis, this list (actually a table) illustrates the most frequent concepts of the loaded input sources, showing the most frequent one at the top. However, this list can be filtered, e.g. by selecting only the common concepts of two specific sources selected from the heat-map matrix (Figure 5.C). Figure 6 shows the result after having selected to view the six common concepts of sources "Common Data Elements Myeloid Leukemia" and "NCT00611247". The filters can be further selected, deselected or adapted based on user-click within the heat-map matrix (Figure 5.C). Following descriptions apply to columns of the table:

- Concept: Concept Code and Concept name according to UMLS
 - Each concept can be furtherly expanded (click on ) to view details of the source items that were coded with concept (Figure 7)
- Concept Rank: Absolute frequency rank of the concept considering all concepts
- Similarity: Summarizes a string-based similarity real value from 0..1 based on Jaccard-Distance of original item attributes (item questions, measurement unit, datatype) coded with this concept. 0 indicate no similarity of the original items, 1 states that all original item attributes were identical.
- All: Absolute concept frequency

Concept	Concept Rank	Similarity	# All	# NCT00512083 ☐ Show only > 0	# Common Data Elements Myeloid Leukemia (MDS, AML, CML) ☒ Show only > 0	# NCT00611247 ☒ Show only > 0	# NCT00512083 (2) ☐ Show only > 0
 C0021430: Informed Consent	1	0,579	3	0	2	1	0
 C1520224: ECOG performance status	3	0,585	3	0	1	1	1
 C0201976: Creatinine measurement, serum (procedure)	7	0,332	2	0	1	1	0
 C0201836: Alanine aminotransferase measurement	11	0,120	2	0	1	1	0
 C0201899: Aspartate aminotransferase measurement	15	0,120	2	0	1	1	0
 C0373595: Creatinine clearance measurement	16	0,379	2	0	1	1	0

Figure 6. Filtered list of most common concepts. Red Rectangle to view source item details (see Figure 7). Green Rectangle: Filter sources.

Step 4) Add items to cart and download resulting item catalogue.

The user may select suitable items in the detailed source item view of the frequency-sorted list, see Figure 6-8. Each item definition can be added to a resulting item catalogue (click Add to cart, Figure 8). To download the item catalogue go to the download option (right upper view, Figure 5). Then, different download formats can be selected, see Figure 9.

	C1520224: ECOG performance status	3	0,585	3	0	1	1	1
--	-----------------------------------	---	-------	---	---	---	---	---

Concept	Count	Question	Type	Data Type	Measurement Unit	Source File	
C1520224: ECOG performance status	1	ECOG performance status finding	ItemDef	text	Null	Common Data Elements Myeloid Leukemia (MDS, AML, CML)	Add to cart
C1520224: ECOG performance status	1	Does the subject have an Eastern Cooperative Oncology Group (ECOG) Performance Status (PS) greater than 2?	ItemDef	boolean	Null	NCT00512083 (2)	Add to cart
C1520224: ECOG performance status	1	6. ecog performance status greater than 2.	ItemDef	boolean	Null	NCT00611247	Add to cart

Figure 7. Extended view of the concept “ECOG performance status” to view its source item details. The first source item had the original question “ECOG performance status” and appeared in the source “Common Data Elements Myeloid Leukemia”. Since this item is linked to a codelist (list of permissible values) in the source, it can be additionally expanded (click on) to view the codelist items = permissible values, see Figure 8

Concept	Count	Question	Type	Data Type	Measurement Unit	Source File	
C1520224: ECOG performance status	1	ECOG Performance status finding	ItemDef	integer	Null	ODM Description	Add to cart
• ECOG Performance Status 0: Asymptomatic (Fully active, able to carry on all predisease activities without restriction) • ECOG Performance Status 1: Symptomatic but completely ambulatory (Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature. For example, light housework, office work) • ECOG Performance Status 2: Symptomatic, <50% in bed during the day (Ambulatory and capable of all self care but unable to carry out any work activities. Up and about more than 50% of waking hours) • ECOG Performance Status 3: Symptomatic, >50% in bed, but not bedbound (Capable of only limited self-care, confined to bed or chair 50% or more of waking hours) • ECOG Performance Status 4: Bedbound (Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair) • ECOG Performance Status 5: Death							

Figure 8. Expanded view of the codelist. The user may add this item with all its definitions (codelist, datatype, measurement unit -> not available) to a customer cart of user selected items, which can be download in various standardized format, see Figure 9.

[Back to Analysis](#)

CDEGenerator

Download as ▾

ODM
OpenClinica
REDCap
Macro

Group Name	#Items
Selected Items	1

Question	Concept	Object Type	Data Type	Measurement Unit
ECOG Performance status finding	C1520224	ItemDef	integer	Null

Figure 8. Selected item “Ecog Performance status finding” was added and can now be downloaded as a data model in various research formats.