German Medical Data Sciences 2023 — Science. Close to People.
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI230723

Design and Implementation of an ETL-Process to Transfer Wound-Related Data into a Standardized Common Data Model

Mareike PRZYSUCHA[a1], Jens HÜSERS[a], Daniil LIBERMAN[b], Oliver KERSTEN[b], Aphrodite SCHLÜTER[b], Sebastian FRAAS[b], Dorothee BUSCH[c], Maurice MOELLEKEN[d], Cornelia ERFURT-BERGE[c], Joachim DISSEMOND[d], and Ursula HÜBNER[a]

[a] *Health Informatics Research Group, Osnabrück University of AS, Germany*
[b] *apenio GmbH & Co. KG, Germany*
[c] *Department of Dermatology, Uniklinikum Erlangen, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Erlangen, Germany*
[d] *Department of Dermatology, Venerology and Allergology, University Hospital of Essen, Germany*

Abstract. For observational studies, which are relevant especially for chronic conditions like chronic wounds, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) offers a standardized database schema. In this study an ETL process for the transition of wound related data was developed. After understanding the data in general and mapping the relevant codes to concepts available in OMOP, the ETL process was implemented. In a first step, a generic algorithm to convert data to a csv format was implemented in Java. The resulting csv file was then processed within KNIME to be loaded into an OMOP CDM conformant database. During the whole ETL process, HL7 FHIR CodeSystem and ConceptMap resources were used for coding and mapping. First clinical test cases to retrieve data were successfully processed as an example to demonstrate the feasibility and usefulness. They concerned wound size at the first visit and the main issues of patients in the wound quality of life questionnaire (n = 24). In general, the ETL process worked well, yet some challenges arose, like post coordinated SNOMED codes or conditions, which might occur more than once.

Keywords. OMOP CDM, ETL Process, Wound Care

1. Introduction

To evaluate the outcome and impact of health related procedures for patients with chronic conditions, studies based on observational data are a promising approach next to randomized trials [1]. To enhance observational studies, the Observational Medical Outcomes Partnership (OMOP) was formed in 2008. They created a common data model (CDM) allowing researchers to perform observational studies with data from different sites and countries. After the end of the partnership in 2013, the Observational Health Data Science and Informatics (OHDSI) initiative was founded in 2014 and continues

[1] Corresponding Author, Mareike Przysucha, Health Informatics Research Group, University AS Osnabrück, Germany; E-mail: m.przysucha@hs-osnabrueck.de

until now. Next to a standardized database schema, the OMOP CDM contains national and international classifications and terminologies like ICD-10-GM [2], SNOMED CT [3], and LOINC [4], but also provides some OMOP specific vocabularies. All terminologies and classifications are mapped to a standard concept to allow researchers to use the same vocabulary when requesting data in national and international studies [5].

One chronic condition putting a high burden on the patients and the health care system are chronic wounds [6]. Though some types are well studied, like the diabetic foot ulcers or arterial or venous leg ulcers, others remain a challenge, like the Pyoderma gangrenosum [7] or other rare diseases. Observational studies conducted in national or international teams could be a way to learn more about these different wound types.

Therefore, the aims of this study were to investigate the applicability of the OMOP CDM in the context of wound-related data, to build a pertaining ETL process, to test the result based on clinical test cases and hereby identify challenges that arose during this process.

2. Methods

To answer the research question, an ETL process was developed according to recommendations of the OHDSI community [5].

2.1 Understanding the data

This study was based on a standardized data set, the national consensus for documenting leg ulcers [8,9], which was agreed by the clinicians in this study to be applicable for other wound types as well. This data set was implemented in the software apenio® by the vendor. The software was extended with an interface to deliver the data in an eXtensible Markup Language (XML) format. Next to information about the patient and the wound (mainly the patient number and wound ID), the data fields comprised among others the following elements: (a) a unique name of the data field, e.g., wound size, (b) the creation date and time, e.g., 01.02.2023 12:07, (c) the creator name, (d) the update date and time, e.g., 01.02.2023 12:10, (e) the name of the updater, and (f) the value represented as string, e.g., 3x4 for the wound size.

For each information represented in such a manner, a health IT specialist manually performed an analysis of the data for (a) the real data type (as data were transmitted as text) and (b) the possible values in case the data came from a limited set. Based on this information, a Unified Modelling Language (UML) class diagram was created, supporting the identification of information coherent in content and the description of their relationship. Afterwards, the same health IT specialist mapped the data to the OMOP CDM V5.4 tables and columns and documented these mappings.

2.2 Code mappings

Based on the analysis in step 2.1, customized code systems were created in HL7 FHIR representing the codes for coded content. These code systems were mapped to SNOMED CT, using pre-coordinated codes only, based on a previous study [10]. The mappings were represented as HL7 FHIR ConceptMaps. Also, all data requiring a code

were assigned either a SNOMED CT or a LOINC code whenever possible while preferring SNOMED CT codes. Unmapped codes were also documented.

2.3 Implementation of the ETL process

The ETL process was split into two sub ETL processes.In the first process, the XML output from the apenio® documentation software was read into a Java Servlet transforming it into a csv file together with a SPSS syntax file to allow clinicians to work on the data themselves. The transformation was designed quite generic, allowing information to be included or removed with little changes. This generic approach required (a) HL7 FHIR CodeSystem resources for each coded content and Booleans, and (b) a csv file containing a list of all columns in the final csv file with:

- the name of the data field in the XML export,
- the column number of the data in the resulting csv file,
- the data type in SPSS for the SPSS syntax file,
- the name of the CodeSystem (a) for the transformation from string to code as well as (b) for the generation of the SPSS syntax file to add value labels so that clinicians can understand the data, and
- the name of the variable in SPSS / the csv file.

The Java program was deployed on a Java Server and a graphical user interface was provided so that clinicians could obtain the csv file and SPSS syntax file.

In the second sub-process, an ETL process was implemented to transfer the data coming from the csv file into an OMOP CDM conformant database. The process was implemented in KNIME [11]. After reading the file into a table, the table was split according to the content: patient demographics, general patient data, confounding conditions, general assessment, medication anamnesis, non-changing wound parameters, changing wound parameters, diagnostics, therapy, treatment goals and patient reported outcome measures.

In the transformation process, the HL7 FHIR ConceptMaps were used to map coded data to SNOMED CT. Codes from SNOMED CT and LOINC were added where applicable, e. g., for the code identifying the type of observation, or for identifying the drug the patient was exposed to. For all LOINC and SNOMED CT codes the OMOP CDM Concept IDs were then extracted, based on a table containing an extract from the Automated Terminology Harmonization, Extraction, and Normalization for Analytics (ATHENA) tool [12], the reference database for OMOP concepts. Standard concept IDs were also extracted from the data base and included into the data set. For six continuous information items, like for the BMI, apenio® represented these values as discrete codes, like '< 16', i.e., 'underweight'. The codes were not mapped as they represented numbers or intervals, instead they were translated to numerical values. Afterwards, unit and/or qualifier codes (like "=" or "≥") were added if necessary. In addition to these transformations, all data were split, so that each information was represented in a single row. Content-related information was first written in one line and only later separated. They were provided with globally unique identifiers (GUID) that allow later association, and information about the nature of the connection was also inserted in both directions. The connection of the items was later converted into the format for the fact_relationship table of the OMOP CDM using the GUIDs and stored in the corresponding table.

Finally, the data sets were prepared for inserting them into the OMOP CDM database. A list with the OMOP tables considered is provided in Table 1.

In the load process, the processed data were written into the OMOP CDM database. The correct sequence was kept in order to ensure referential integrity: location, care_site > person > observation, measurement, condition_occurrence, procedure_occurrence, drug_exposure, device_exposure > fact_relationship.

2.4 Implementation testing

To probe whether the implemented pipeline works, two realistic and relevant test cases were defined together with the clinicians in this study. These test cases considered the main statistics (minimum, maximum, mean, standard deviation, median, and quartiles) for the wound area at the first visit recorded as the wound area is a relevant indicator for wound (healing) status, and the three most urgent points of the patients in the Wound Quality of Life (Wound-QoL) questionnaire as relevant patient reported outcomes. Both test cases were implemented on the OMOP CDM conformant database.

3. Results

3.1 Data Mapping

An overview of the number of fields mapped to the OMOP CDM table is provided in Table 1. In total, 281 information items could be mapped and three information items could not be mapped directly. These items were the wound id, the date, and the composition of wound length and width. While the first were taken and mapped for each information item, the combination of width and length of the wound, expressed as *length* x *width*, was split into width and length, which could be mapped.

Table 1: Overview of total number of fields mapped to the different OMOP tables

OMOP table	Number of fields mapped (+ additional mappings)
care_site	1
condition_occurrence	52
device_exposure	19
drug_exposure	24
fact_relationship	24
location	1
measurement	11
observation	127 (+2)
person	3
procedure_occurrence	19 (+1)
unmapped	3

3.2 Code Mappings

The results of the code mapping are represented in Table 2. In total 415 codes had to be mapped. Out of these, 285 codes were mapped to SNOMED CT concepts (275 pre-coordinated, 10 post-coordinated), 7 codes to LOINC codes, and 4 to OMOP CDM

concepts (gender concepts), while 119 codes remained unmapped (29 unmapped when mapping was unnecessary [intentionally unmapped], 90 unmapped when suitable codes were missing [unintentionally unmapped]).

An analysis of unintentionally unmapped codes showed some reasons for codes remaining unmapped. One reason was that the source concept, such as "pain management according to WHO stage I", was on a much higher aggregation level than SNOMED CT or LOINC. Vice versa, source concepts with a lower aggregation level did not impose such a problem, as a suitable, yet broader code could be found, but with a loss of precision. Another reason for unmapped codes was, that wound specific assessment scales were used (e.g., PARACELSUS score [13], Wound Quality of Life Questionnaire [14,15]), which were not (yet) represented in SNOMED CT or LOINC. In one case, a SNOMED CT code was found, but could not be used, as the code was not available in ATHENA.

It also occurred that concepts in the standardized wound data set could not be mapped to a pre-coordinated, but a post-coordinated SNOMED CT code, like "wound pain when changing the wound dressing" or "hyperchloride acids". These codes needed a workaround as the OMOP CDM does not allow the usage of post-coordinated codes.

Table 2: Overview of code mapping.

Intentionally unmapped = codes for which no mapping was needed; Unintentionally unmapped = codes for which a code was needed, but no appropriate code could be found

Code Mappings	**Unmapped**		**SNOMED CT**		**LOINC**	**OMOP CDM**
	inten-tional	uninten-tional	pre-coordi-nated	post-coordi-nated		
Demographics	10	0	2	0	0	4
Patient data	7	3	24	0	2	0
Conditions	1	3	66	2	1	0
Assessment	0	6	11	0	0	0
Medication	1	4	4	0	0	0
Non-changing	0	5	23	0	2	0
Changing	1	3	64	0	2	0
Diagnostics	0	16	7	2	0	0
Therapy	5	18	66	6	0	0
Goals	1	14	2	0	0	0
PROMs	0	18	6	0	0	0
Other	3	0	0	0	0	0
SUM	**29**	**90**	**275**	**10**	**7**	**4**

3.3 Implementation of the ETL

Several components were set up and implemented to realise this concept. For organisational reasons, the documentation software apenio® was implemented at Osnabrück University of Applied Sciences and made available remotely for the clinicians. A tomcat server was installed, containing a webservice which converted data from XML to csv to be uploaded to the clinicians' devices. The PostgreSQL database for

the OMOP CDM database ran on the same server, while KNIME was run locally due to organisational reasons.

During implementation several challenges arose. One challenge was the possible multiplicity of wound conditions if a patient had several wounds due to the same condition. As the condition_occurrence table could not recognize that there were several wounds due to one condition, several occurrences of the same condition were created.

To avoid this error, a new column `description` was added. The description allowed more information about the type and location of the wound and thus made it possible to uniquely identify a wound.

Another challenge was to represent clinical conditions that were reported several times for a patient, like diabetes, but were present the whole time and therefore recorded for every single visit yet should only be recorded once. Therefore, entries for the condition_occurrences table were reduced to one entry per condition, the start date being the earliest start date the condition was recorded.

A third challenge was, that post-coordinated SNOMED CT codes cannot be directly mapped, as the OMOP CDM only allows the storage of pre-coordinated SNOMED CT codes. Therefore, the intended post-coordination had to be split into several data sets. One example showing this is “wound dressing with antiseptic gel”. In the original data, the key was “primary dressing” or “secondary dressing”, the value was “antiseptic gel”. The mapping and the relationships are shown in Figure 1.

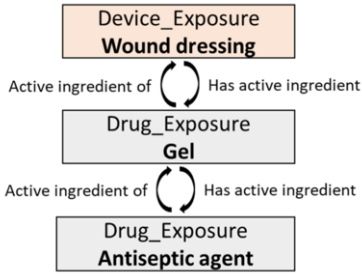

Figure 1: Mapping of "change wound dressing with an antiseptic gel”. The type of relationship between the different data is attached to the arrow.

3.4 Implementation testing

During the transformation process, information about 24 patients were successfully loaded into the database. For both test cases, workflows in KNIME were built, resulting in the relevant information.

The wound area at the time of the first visit was available for n=18 patients. The statistical values computed in KNIME based on the data resulting from the ETL process are shown in Table 3.

Table 3: Results for the wound area in cm² (n=18)

n	min	Q_1	median	Q_3	Max	mean	std
18	0.00	2.00	5.50	23.60	300.00	34.33	73.97

For the Wound-QoL questionnaire, each item could have a value of 0 (not at all), 1 (a little), 2 (moderately), 3 (quite a lot) and 4 (very much). The answers of the 24 patients for the three items with the highest ratings are shown in Table 4.

Table 4: Results for the three items with the highest ratings of the Wound-QoL (n=24)

Item: "In the last seven days ..."	Median
I felt frustrated because the wound is taking so long to heal	Quite a lot
I was worried about my wound	Between moderately and quite a lot
I was afraid of the wound getting worse or of new wounds appearing	Moderately

4. Discussion

In general, OMOP was found to be an appropriate tool to represent wound related data, however, there are some drawbacks. The coverage rate of wound related information in a standard terminology is not complete as was already found in a previous study [10]. Unmapped codes had to be either written in free text or post-coordination needs to be split like shown above. Another drawback was the handling of conditions, which might occur more than once and cannot be totally distinguishable, like wounds. Also, treatment goals are not considered in the OMOP CDM, though it is interesting to know what the goals were.

The KNIME tool proved to be a good instrument to design and execute ETL processes. In this context, HL7 FHIR ConceptMaps turned out to be a flexible tool to design mappings with custom systems, allowing the easy addition of codes.

One limitation of this study is that a structured quality assessment did not take place. For this, the Data Quality Dashboard from the OHDSI tool stack [16] could be used as well as ACHILLES for characterization and visualisation of the database or ATLAS for designing and executing analyses.

Another drawback of this study is, that an interface with a proprietary format was used. In the future, HL7 FHIR as an interface to enhance interoperability could be considered. In this context, the applicability of existing FHIR to OMOP implementations, like from the German Medical Informatics Initiative [17] could be checked and developed further in accordance with current national developments like ISiK (information systems in hospitals) [18], ISiP (information systems in nursing) [19], or MIOs (medical information objects) [20].

This study presents an example of how an OMOP database could be populated from a documentation software through a defined ETL process. It also demonstrates the clinical usefulness based on clinical sample questions. Eventually, the findings from this study can serve as a leverage for analysing observational data in a clinically highly relevant field that would have been difficult to do otherwise.

5. Conclusion

The ETL process for transferring wound related data into the OMOP CDM could be successfully developed and implemented. However, some challenges arose during the process and several items could not be mapped to a concept in the OMOP CDM, resulting in a loss of interoperability. This limits the ability to conduct cross-language studies because there is no guarantee that exactly the same terms will be used in the textual description in the OMOP CDM, so that data sets may not be found.

Acknowledgement

This project (ZIEL) was funded by the German Federal Ministry of Education and Research (BMBF) (grant: 16SV8616).

Declarations

A positive vote was obtained from the Ethics Commission of the University AS Osnabrück for the data gathering and processing (voteID: HSOS/2021/1/5)

References

[1] R.A. Rosati, K.L. Lee, R.M. Califf, D.B. Pryor, and F.E. Harrell, Problems and advantages of an observational data base approach to evaluating the effect of therapy on outcome. Circulation **65** (1982), 27–32.
[2] ICD-10-GM: Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, German Modification, https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html [cited 2022 September 22].
[3] SNOMED International, SNOMED International, http://www.snomed.org/ [cited 2020 November 3].
[4] Regenstrief Institute Inc., LOINC Users' Guide, https://loinc.org/kb/users-guide/ [cited 2023 March 27].
[5] OHDSI, The Book of OHDSI: Observational Health Data Sciences and Informatics, 25.11.2020.
[6] M. Olsson, K. Järbrink, U. Divakar, R. Bajpai, Z. Upton, A. Schmidtchen, and J. Car, The humanistic and economic burden of chronic wounds: A systematic review. Wound Repair Regen **27** (2019), 114–125.
[7] Deutsche Dermatologische Gesellschaft, S1-Leitlinie: Pyoderma gangraenosum: AWMF-Register-Nr.: 013 – 091, 2020, 2020.
[8] K. Herberger, K. Heyer, K. Protz, A. Mayer, J. Dissemond, S. Debus, T. Wild, J. Schmitt, and M. Augustin, Nationaler Konsensus zur Wunddokumentation beim Ulcus cruris: Teil 2: Routineversorgung – Klassifikation der Variablenausprägungen. Hautarzt **68** (2017), 896–911.
[9] K. Heyer, K. Herberger, K. Protz, A. Mayer, J. Dissemond, S. Debus, and M. Augustin, Nationaler Konsensus zur Wunddokumentation beim Ulcus cruris: Teil 1: Routineversorgung – „Standard-Dataset“ und „Minimum-Dataset“. Hautarzt **68** (2017), 740–745.
[10] J. Hüsers, M. Przysucha, M. Esdar, S.M. John, and U.H. Hübner, Expressiveness of an international semantic standard for wound care: Mapping a standardized item set for leg ulcers to the Systematized Nomenclature of Medicine-Clinical Terms. JMIR Med Inform **9** (2021), e31980.
[11] KNIME AG, KNIME Analytics Platform.
[12] Odysseus Data Services, Inc., ATHENA.
[13] F. Jockenhöfer, U. Wollina, K.A. Salva, S. Benson, and J. Dissemond, The PARACELSUS score: a novel diagnostic tool for pyoderma gangrenosum. Br J Dermatol **180** (2019), 615–620.

[14] C. Blome, K. Baade, E.S. Debus, P. Price, and M. Augustin, The "Wound-QoL": a short questionnaire measuring quality of life in patients with chronic wounds based on three established disease-specific instruments. Wound Repair Regen **22** (2014), 504–514.
[15] M. Augustin, E. Conde Montero, N. Zander, K. Baade, K. Herberger, E.S. Debus, H. Diener, T. Neubert, and C. Blome, Validity and feasibility of the Wound-QoL questionnaire on health-related quality of life in chronic wounds. Wound Repair Regen **25** (2017), 852–857.
[16] OHDSI, Software Tools, https://www.ohdsi.org/software-tools/ [cited 2023 May 3].
[17] Y. Peng, E. Henke, I. Reinecke, M. Zoch, M. Sedlmayr, and F. Bathelt, An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform **169** (2023), 104925.
[18] gematik GmbH, Interoperabilität dank ISiK, https://fachportal.gematik.de/informationen-fuer/isik [cited 2023 May 3].
[19] gematik GmbH, Pflege & TI, https://fachportal.gematik.de/informationen-fuer/isip [cited 2023 May 3].
[20] MIO: Medizinische Informationsobjekte, https://mio.kbv.de/site/mio# [cited 2022 May 11]