German Medical Data Sciences 2021: Digital Medicine: Recognize - Understand - Heal
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210549

Hands on the Medical Informatics Initiative Core Data Set — Lessons Learned from Converting the MIMIC-IV

Hannes ULRICH[a,1], Paul BEHREND[a], Joshua WIEDEKOPF[a], Cora DRENKHAHN[a], Ann-Kristin KOCK-SCHOPPENHAUER[a] and Josef INGENERF[a,b]

[a] *IT Center for Clinical Research (ITCR-L), University of Lübeck, Lübeck, Germany*
[b] *Institute of Medical Informatics, University of Lübeck, Lübeck, Germany*

Abstract. With the steady increase in the connectivity of the healthcare system, new requirements and challenges are emerging. In addition to the seamless exchange of data between service providers on a national level, the local legacy data must also meet the new requirements. For this purpose, the applications used must be tested securely and sufficiently. However, the availability of suitable and realistic test data is not always given. Therefore, this study deals with the creation of test data based on real electronic health record data provided by the Medical Information Mart for Intensive Care (MIMIC-IV) database. In addition to converting the data to the current FHIR R4, conversion to the core data sets of the German Medical Informatics Initiative was also presented and made available. The test data was generated to simulate a legacy data transfer. Moreover, four different FHIR servers were tested for performance. This study is the first step toward comparable test scenarios around shared datasets and promotes comparability among providers on a national level.

Keywords. Medical Informatics, Data Integration, Information Dissemination, MIMIC-IV, HL7 FHIR

1. Introduction

1.1. Background

With the steady increase of applications in the digitized healthcare system, the need for data integration is growing in order to ensure the exchange of healthcare data between providers on the one hand [1] and to ensure the continued use of legacy data when it comes to an application replacement on the other hand [2]. Data exchange on a national level remains a challenge, which is why the German Medical Informatics Initiative (MII) has set this as the main goal [3]. In this initiative, all kinds of organizational, legal, or ethical tasks need to be solved; furthermore, the technical provision of patient data and interoperable communication are significant hurdles. For this reason, four consortia are collaborating to develop a harmonized core data set (MII-CDS) as standard schemas for

[1] Hannes Ulrich, IT Center for Clinical Research, Lübeck; Telephone: +49 (0) 451 - 3101 5607; Fax: +49 451 3101 5604; E-mail: h.ulrich@uni-luebeck.de

the communication of all participating sites [4]. The MII-CDS is developed using HL7 Fast Healthcare Interoperability Resources (FHIR) [5], a emerging standard for the exchange of healthcare-related information. In particular, the MII-CDS modules are modeled as FHIR resources and constraining profiles and frequently published [6]. The MII-compliant export of sensitive patient data is performed by local data integration centers, which must complete the difficult task of mapping and migrating their local data schemas to the harmonized MII-CDS. Ensuring secure communication of sensitive data requires precautions and sufficient testing in advance. However, valid test data that is as close to the real data as possible is often lacking because of legal requirements or the early development and set-up phases of the data integration centers. But yet, having usable test data is essential and creates comparability on the national level across all sites. There are two viable scenarios to obtain such a test dataset: a synthetic generation [7] or the use of a public, anonymized dataset. Some MII-CDS-compliant, synthetic data sets are available for the cross-consortium use case CORD-MI. This is a valuable first step, but not all MII-CDS modules are covered, and variation of the patient data is minimal, severely compromising its usability. In our approach, we transformed the Medical Information Mart for Intensive Care (MIMIC) to the MII-CDS. The MIMIC-IV [8,9] is a database of clinical patient data from primarily intensive care units admitted to the Beth Israel Deaconess Medical Center in the United States. The data is de-identified and ranges from 2008 to 2019. The data set includes 383,000 unique patients with 524,000 admissions, 122 million laboratory events, and 17 million drug prescriptions, along with related data tables. The data is freely available to any credentialed user of the PhysioNet, a repository of freely available medical research data and can be used for scientific purposes. In previous studies, we addressed the direct transformation from previous MIMIC versions into HL7 FHIR and gained insights into the technical limitations of FHIR servers processing a large amount of data at once (bulk imports) [10,11].

1.2. Objective and Requirements

This study investigates legacy data integration by using the recently adopted MII-CDS for ETL processes to data warehouses. The aim is to evaluate the transformation of MIMIC-IV into the current release FHIR R4 for general use and to implement all available MII-CDS modules for testing data exchange in Germany. However, it must first be analyzed whether the MIMIC data set provides all the data required by the modules to identify pitfalls or inconsistencies. A significant challenge is the consistent use of the terminologies proposed by the MII-CDS. Since the MIMIC-IV originates from the US-American healthcare system, the German MII-CDS conformant representation will require adaptions and terminology mappings to the mandatory coding systems.

2. State of the art

2.1. Related Work

Many approaches use MIMIC to test newly developed methods since it is unparalleled in its status as a freely available dataset containing rich and real patient data. Noteworthy is the study of Roehrs et al.[12] describing the integration of HL7 FHIR and MIMIC-III together in a proposed higher-level scheme and the study of Wang et al.[13] presenting an open-source extraction pipeline to ease the preprocessing for machine-learning methods. In a literature search regarding the usage of the MIMIC-IV database, no paper could be found that targeted the conversion of the dataset into standardized representations

2.2. Shortcomings

There is currently no other approach that transforms the newly released MIMIC-IV data mart into the newest release FHIR R4. Furthermore, to our knowledge, there is no existing work that addresses MII-CDS-compliant transformation of MIMIC-IV or other legacy data transfer.

3. Concept

At the beginning of the conceptual design of the data conversion, it was considered which data of the database should be converted and transferred to which FHIR resources and which profiles. The relations in the MIMIC database do not necessarily correspond to FHIR resources, and likewise, not all data can be mapped unambiguously.

The MIMIC database consists of a total of 27 tables. The core of the database is the table *admissions*, which represents all hospitalizations. Linked to this table are the corresponding patients with deidentified demographic data (*patients*), the intensive care unit stays (*icustays*), and the various events, observations, and prescriptions that occurred during a stay (*chartevents, diagnoses_icd, labevents, procedures_icd, noteevents, prescriptions,* and others). In addition, there are tables for assigning codes and in-house transfers. The conversion to FHIR resources resulted in a similar scheme: Patients from the corresponding table are transferred to *Patient* resources. A hospitalization from admissions corresponds to a FHIR *Encounter*, referencing the respective patient and further observations and results.

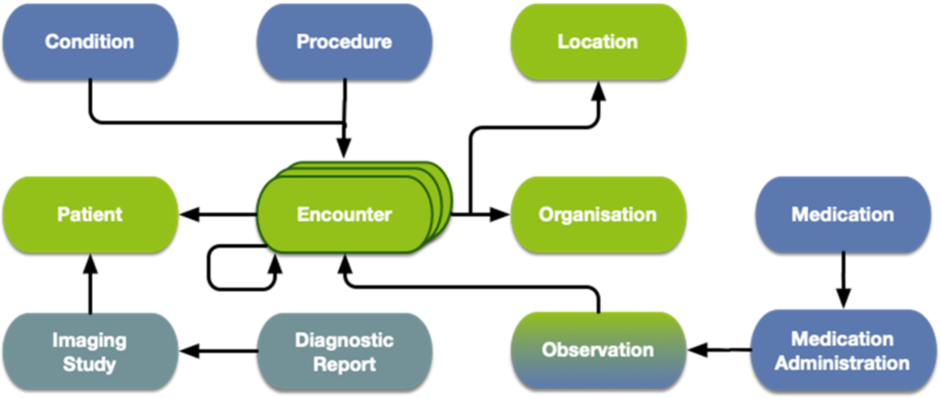

figure 1 Schematic overview of all used, interconnected FHIR resources to map the MIMIC IV data structures. The central resource is the Encounter which describes the entire hospitalization with all in-house transfers and connects every resource together. The colorization indicates the affiliation to the MII-CDS modules, adapted from the current release [14]. Green are the base modules, blue the extension modules and petrol the added standard resources.

table 1 The table shows the used MIMIC-IV major tables and the resulting HL7 FHIR resources.

MIMIC IV tables	FHIR Resource	CDS-Profile	Comment
core.patients	**Patient**	**Patient**	**BirthDate not possible, because such data is masked in MIMIC; Placeholder Address, because MIMIC does not contain it, but it is required in the profile.**
core.admissions	**Encounter**	**KontaktGesundheitseinrichtung**	
core.transfer	**Organization** **Location**		**We used the encounter, organization and location for the transfer chain within the hospital.**
icu.chartevents + icu.d_items	**Observation**	**Vitalstatus**	**We used the LOINC magic numbers from the Observation Vital Signs profile to determine if it was a vital sign.**
hosp.ICD_Diagnosis	**Condition**	**Diagnose**	
hosp.procedure_icd	**Procedure**	**Procedure**	
hosp.prescriptions	**Medication**	**Medication**	**Components are not available for all medications: in this case, only indication of absence of data.**
	MedicationAdministration	**MedicationAdministration**	**Start and end date not always available (e.g. in case of input errors in MIMIC-DB): in this case only indication of the absence of the data.**
hosp.labevent	**Observation**	**ObservationLab**	
CXR.studies	**ImagingStudy**		**No matching KDS resource at the time of the project; use of the default FHIR profile.**
CXR.records	**DiagnosticReport**		**No matching KDS resource at the time of the project; use of the default FHIR profile.**

With the release of the new MIMIC-IV, the table structure has also been adapted and simplified. Some information is even completely omitted in this regard. In contrast to the previous MIMIC version, *caregivers* are no longer included, so there are unfortunately no corresponding practitioners. Furthermore, clinical notes and waveform data from the ICU are not published yet. A new data set (*CXR*) for chest radiographs and the corresponding reports has been published, which are modeled as a *DiagnosticReport* and the associated *ImagingStudies,* as seen in table 1. From a terminological point of view, in addition to ICD-9 Clinical Modification Diagnosis and Procedure codes, the successors ICD-10 Clinical Modification and ICD-10 Procedure Coding System are now being used. In addition, the controlled terminology LOINC is used for the laboratory tests and the *National Drug Code* (NDC) or the *Generic Sequence Number* (GSN) for drugs. The MII-CDS consists of six basic modules and a large number of extension modules in the current release [14]. The basic modules are defined in a cross-disciplinary context, whereas the extension modules cover specific applications and areas of expertise. Our study covers the six basic modules: *Person, Encounter, Diagnosis, Procedure, Laboratory Results*, and *Medication*. Subsequently, we anticipated the imaging extension module to represent the *CXR* chest x-ray data, as seen in Figure 1. Regarding the module *Encounter*, a differentiation was made according to the MII-CDS facility or department contact. The postponed billing case and the optional supply point contact could not be taken into account due to the absence of that information.

In reviewing the modules, it was determined that the following terminologies need to be used and mapped to be compliant with the German national requirements: ICD-10 German Modification for coding diagnoses, Operation and Procedure Classification System (OPS) for coding procedures, SNOMED CT for coding procedures and diagnosis, Anatomical Therapeutic Chemical Classification System (ATC) for coding the contents of pharmaceutical products, and Pharmacy Central Number (PZN) for identifying medicinal products available on the German market. Since there was no overlap between the terminologies given in the MIMIC-IV and the coding systems required by the MII-CDS, suitable mappings had to be created. The MII-CDS extension module "Intensive Care" was not available at the time of the study. The measurements were mapped according to the CDS Vitalstatus profile since most of the data originated from vital monitoring. A separate additional module would nevertheless be helpful.

4. Implementation

4.1. Solution/Results

We propose a program called MIMIC4FHIR that converts the MIMIC-IV into the native FHIR R4 specification, and addtionally to the profiles within the German Core data sets. In a previous study, the processing times of the FHIR servers were shown to be problematic performance-wise [11]. Aware of the issue, the program is separated into two parts: transformation and transaction. The information transformation is done in a case-by-case process for each patient resulting in a single FHIR transaction bundle per case. In addition, a bundle is created for the imaging-related CXR data. The single resource entries are defined as *conditional creates* to avoid redundant data collection within the FHIR repository. The MII-CDS mandates multiple coding systems, so we had to create suitable, machine-processable mappings for the transformation. For this, we used OHDSI Athena, which maintains and provides a variety of terminologies and curated mappings [15]. Through appropriate data fusions, FHIR ConceptMaps were created for the transformation of ICD-9 and ICD-10 into ICD-10-GM as required by MII-CDS. Furthermore, a ConceptMap of ICD-9 Procedure to SNOMED CT was created, while a mapping of OPS was unfortunately not possible due to lacking data availability. The created ConceptMaps were stored in a CSIRO Ontoserver, a specialized FHIR server implementation of the FHIR Terminology Module [16]. The RxNorm API from the National Institutes of Health was used for the MII-CDS-compliant representation of medication data. This API allows the lookup of NDC or GSN codes in the RxNorm data, where ATC information used for coding is available. If no suitable mapping was available, the MII-CDS-compliant FHIR extension *data-absent-reason* was used. To check and ensure validity, the built-in FHIR Instance Validator was augmented with the MII-CDS StructureDefinitions and applied to each resource.

RabbitMQ, an implementation of the *Advanced Message Queuing Protocol* [17], was used for the transaction, i.e., the output of the converted FHIR bundles. The bundles are written to a queue, and a consumer thus processes the bundles independently of the data transformation. There are different derivation options provided: saving the output as XML files or direct transmission to a FHIR server.

4.2. System in Use

MIMIC4FHIR was developed for test data generation and is also being further refined for this purpose. The source code is freely available[2] and can help other MII sites in the further development of their integration scenarios. The ongoing changes to the MII-CDS are constantly integrated into the program to guarantee continuous availability of test data. Several performance improvements have been made during the development mitigating problems we faced in the earlier studies. These were, on the one hand, on the database side, e.g., more indexes for faster queries, or, on the other hand, multi-thread support to create ten patients simultaneously. An additional optimization step included the caching of requests to the external knowledge sources, RxNorm API and the Ontoserver, to reduce processing time and network traffic. At the current stage of development, the creation of a patient takes about one second on average, while some patients may require significantly more time if more data is available for that patient. This is a significant performance increase compared to the 180 seconds per patient in the previous work. In order to properly evaluate a legacy data transfer and a bulk import with the generated data, comparisons were made with four different FHIR servers: HAPI FHIR [19], Blaze [20], IBM FHIR Server [21], and Firely Server [22]. The four servers were run on Photon OS VMs using HDDs with two different configurations: 2 CPU and 8 GB RAM or 8 CPU with 32 GB RAM. The applications were run with default configurations over Docker. For the evaluation, test data with 100, 1,000 and 10,000 patients were previously generated and transferred to the respective servers. We observed the processing time and the performance gain with more CPU core and RAM — the results are shown in table 2.

Table 2 The table shows the results of the data throughput tests. Each server was tested three times with two different hardware configurations.

Server /	Setting 2 CPU and 8 GB RAM			Setting 8 CPU and 32 GB RAM		
Patients	100	1,000	10,000	100	1,000	10,000
HAPI FHIR	0:00:28	0:03:58	0:15:04	0:00:18	0:04:03	0:16:06
Blaze	0:00:02	0:00:35	0:01:24	0:00:02	0:00:13	0:03:36
IBM FHIR	0:04:50	0:47:31	2:52:49	0:04:47	0:47:26	2:48:12
Vonk	0:00:03	0:00:42	0:20:47	0:00:02	0:04:24	0:20:42

5. Lessons learned

Three problem areas have been identified in the development of the concept, implementation, and subsequent evaluation: the origin of the data set, the terminologies, and the state of development of the FHIR servers. The MIMIC-IV data set is an important asset to medical informatics research, but problems arise from its U.S. origin and associated coding regarding the use within a German medical informatics setting. Not all information is available to represent every potential use case, e.g., billing data. But an important aspect compared to using synthetic test data is the authenticity of the dataset. It is desirably imperfect: Data is missing or incomplete. For example, patients' dates of birth are masked for anonymization purposes, or the date of a diagnosis is simply not

[2] https://github.com/itcr-uni-luebeck/mimic4fhir

recorded. This allows applications to be tested much more realistically, as they are closer to the real data than synthetic derivatives.

The major problem in the conversion process were the mappings to the required code systems. The MIMIC-IV, due to its U.S. origin, references other vocabularies that had to be mapped. The challenge was, on the one hand, to find suitable source data for creating the mappings, which had the best unique code-to-code associations. On the other hand, the data had to be machine-processable so that it could be transformed into FHIR ConceptMaps. OHDSI Athena [15] was a promising source providing a broad coverage of mappings, but often a direct mapping was not possible. We had to make an intermediate step via a mapping to SNOMED CT. This resulted in some source codes then being mapped to a multitude of target codes, i.e., with no clear code-to-code association. For some codes, no mapping could be identified either, so the extension data-absent-reason was used. In general, the transition between coding systems is a complex challenge, as in result these mappings are difficult to create. This is also the crux: the MIMIC-IV in a suitable format can be used for many other projects. For this, however, the code mapping would have to be validated by experts in order not to jeopardize the conclusion of the projects built on it. This work mainly dealt with technical aspects such as the provision of test data and therefore does not guarantee mapping correctness.

A major insight relating to the performance of the algorithm relates to the use of caching requests to external knowledge sources, namely the FHIR terminology server and the RxNorm API. Especially important is the caching not only of successful mapping operations, but also of missing mappings. This addition alone enabled a substantial performance gain, as the remote HTTP requests are still a major bottleneck. Also, the use of caching pays off increasingly as more and more patients are processed, since the codes generally seem to follow logarithmic distributions. For example, the most common 50 ICD-9 procedure codes in the table account for over 50% of all ICD-9 procedure codes available. Without caching, those same codes would have to be queried repeatedly. Lastly, the performance and level of development of the servers used vary widely. According to our evaluations, Vonk and Blaze are very performant; HAPI is also acceptable. The IBM server showed a lagging performance compared to the others. In general, it is surprising that performance deteriorates with more resources.

6. Conclusion

Appropriate test data can more securely enable national data exchange testing and thus advance medical informatics research in general. The MIMIC-IV and the introduced conversion to FHIR can contribute their part. The conversion to the German MII-CDS is a first successful step into comparable test scenarios and must continue to be adapted to the new requirements. In addition to being used as test data, the transformed data was also made available for future research. Further projects can now build on the dataset if the mappings used are validated and approved by clinical experts. The used architecture is adaptable for further data sets and shows that a legacy data transfer into the MII-CDS is applicable for staging in data warehouses. With the provision of suitable test data, medical informatics research in general can be advanced. The development of new innovative services can now be better initiated due to the availability.

Declarations

Contributions of the authors: HU, PB and JW developed the approach. CD supported the mapping creation. AKK-S and JI contributed conceptually and conducted review and editing. All authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgement: This work was supported by the HiGHmed project within the Medical Informatics Initiative, funded by the Federal Ministry of Education and Research (grant number 01ZZ1802Z).

References

[1] R.K. Saripalle, Fast Health Interoperability Resources FHIR: Current Status in the Healthcare System, *Int. J. E-Health Med. Commun.* **10** (2019) 76–93. doi:10.4018/IJEHMC.2019010105.

[2] S.M. Meystre, C. Lovis, T. Bürkle, A. Budrionis, and C.U. Lehmann, Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress, *Yearb Med Inform.* **26** (2017) 38–52. doi:10.15265/IY-2017-007.

[3] S.C. Semler, F. Wissing, and R. Heyder, German Medical Informatics Initiative, *Methods Inf Med.* **57** (2018) e50–e56. doi:10.3414/ME18-03-0003.

[4] T. Ganslandt, M. Boeker, M. Löbe, F. Prasser, J. Schepers, S. Semler, S. Thun, and U. Sax, Der Kerndatensatz der Medizininformatik-Initiative: Ein Schritt zur Sekundärnutzung von Versorgungsdaten auf nationaler Ebene, *Forum Der Medizin-Dokumentation Und Medizin-Informatik.* **20** (2018) 17–21.

[5] T. Benson, and G. Grieve, Principles of Health Interoperability - SNOMED CT, HL7 and FHIR, 3rd ed., Springer International Publishing, London, 2016. doi:10.1007/978-3-319-30370-3.

[6] HL7 FHIR, SIMPLIFIER.NET - Medizininformatik Initiative (MII), (2021). https://simplifier.net/organization/koordinationsstellemii/~home (accessed March 11, 2021).

[7] J. Wiedekopf, H. Ulrich, A. Essenwanger, A. Kiel, A.-K. Kock-Schoppenhauer, and J. Ingenerf, Desiderata for a Synthetic Clinical Data Generator - Requirements from an International Perspective, in: Studies in Health Technology and Informatics, In Press, 2021.

[8] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, and H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation.* **101** (2000) e215–e220.

[9] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L.A. Celi, and M. Roger, MIMIC-IV (version 1.0), (2021). https://doi.org/10.13026/a3wn-hq05. (accessed January 12, 2021).

[10] C. Kamann, and J. Ingenerf, Transformation von Intensivdaten von der MIMIC-II Datenbank auf einen FHIR-Server, in: HEC 2016: Health – Exploring Complexity. Joint Conference of GMDS, DGEpi, IEA-EEF, EFMI., German Medical Science GMS Publishing House, München, 2016.

[11] S. Ververs, H. Ulrich, A.-K. Kock, and J. Ingenerf, Konvertierung von MIMIC-III-Daten zu FHIR, *63. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS).* (2018). doi:10.3205/18gmds018.

[12] A. Roehrs, C.A. da Costa, R. da Rosa Righi, S.J. Rigo, and M.H. Wichman, Toward a model for personal record interoperability, *IEEE Journal of Biomedical and Health Informatics.* **23** (2018) 867–873.

[13] S. Wang, M.B. McDermott, G. Chauhan, M. Ghassemi, M.C. Hughes, and T. Naumann, Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii, in: 2020: pp. 222–235.

[14] TMF e.V. (Geschäftsstelle), Der Kerndatensatz der Medizininformatik-Initiative | Medizininformatik-Initiative, (2020). https://www.medizininformatik-initiative.de/de/der-kerndatensatz-der-medizininformatik-initiative (accessed March 19, 2021).

[15] C. Maier, L. Lang, H. Storf, P. Vormstein, R. Bieber, J. Bernarding, T. Herrmann, C. Haverkamp, P. Horki, and J. Laufer, Towards implementation of OMOP in a German university hospital consortium, *Applied Clinical Informatics.* **9** (2018) 54.

[16] A. Metke-Jimenez, J. Steel, D. Hansen, and M. Lawley, Ontoserver: a syndicated terminology server, *Journal of Biomedical Semantics.* **9** (2018). doi:10.1186/s13326-018-0191-z.

[17] S. Vinoski, Advanced message queuing protocol, *IEEE Internet Computing.* **10** (2006) 87–89.

[19] Smile CDR, HAPI FHIR, (2021). https://hapifhir.io/ (accessed July 26, 2021).

[20] samply/blaze, (2021). https://github.com/samply/blaze (accessed July 26, 2021).

[21] IBM FHIR Server, (2021). https://ibm.github.io/FHIR/ (accessed July 26, 2021).

[22] FHIR Server | Flexible and Reliable FHIR Server for Health Organizations, *Firely.* (2021). https://fire.ly/products/firely-server/ (accessed July 26, 2021).