

Synthetic Data Generation in Hematology - Paving the Way for OMOP and FHIR Integration

Waldemar HAHN ^{a,b,1}, Najia AHMADI ^b, Katja HOFFMANN ^b,
Jan-Niklas ECKARDT ^{c,d}, Martin SEDLMAYR ^b and Markus WOLFIEN ^{b,a}
^a Center for Scalable Data Analytics and Artificial Intelligence, Dresden, Germany
^b Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav
Carus, Technische Universität Dresden, Dresden, Germany
^c Department of Internal Medicine I, University Hospital Carl Gustav Carus,
Technische Universität, Dresden, Germany
^d Else Kröner Fresenius Center for Digital Health, Technische Universität Dresden,
Germany

ORCID ID: Waldemar Hahn <https://orcid.org/0009-0007-7934-930X>,

Markus Wolfien <https://orcid.org/0000-0002-1887-4772>

Abstract. This study advances the utility of synthetic study data in hematology, particularly for Acute Myeloid Leukemia (AML), by facilitating its integration into healthcare systems and research platforms through standardization into the Observational Medical Outcomes Partnership (OMOP) and Fast Healthcare Interoperability Resources (FHIR) formats. In our previous work, we addressed the need for high-quality patient data and used CTAB-GAN+ and Normalizing Flow (NFlow) to synthesize data from 1606 patients across four multicenter AML clinical trials. We published the generated synthetic cohorts, that accurately replicate the distributions of key demographic, laboratory, molecular, and cytogenetic variables, alongside patient outcomes, demonstrating high fidelity and usability. The conversion to the OMOP format opens avenues for comparative observational multi-center research by enabling seamless combination with related OMOP datasets, thereby broadening the scope of AML research. Similarly, standardization into FHIR facilitates further developments of applications, e.g. via the SMART-on-FHIR platform, offering realistic test data. This effort aims to foster a more collaborative research environment and facilitate the development of innovative tools and applications in AML care and research.

Keywords. Synthetic patient data, Data Sharing, Hematology, AML, OMOP, FHIR

1. Introduction

In Health Informatics, sharing realistic medical datasets is essential for developing and applying computational tools and visual analytics effectively in healthcare. Synthetic data that accurately reflects real-world clinical data is extremely valuable [1]. It provides essential insights for both those creating health informatics tools and healthcare professionals using them [2]. However, to maximize their impact on clinical practice,

¹ Corresponding Author: Waldemar Hahn; E-mail: waldemar.hahn@tu-dresden.de

synthetic data needs to be integrated into broader, harmonized data formats. Common standards like the Observational Medical Outcomes Partnership (OMOP) and Fast Healthcare Interoperability Resources (FHIR) play a crucial role in ensuring interoperability across healthcare systems and technologies, thus supporting both patient care and research [3]. OMOP organizes clinical data to support observational research, while FHIR facilitates the exchange of data between healthcare systems.

The conversion of the MIMIC-IV dataset to the FHIR framework, and the development of minimal Common Oncology Data Elements (mCODE), as a specific representation of FHIR resources for cancer patients, serve as prominent examples of these efforts [4,5]. Furthermore, the necessity of extending these developments to specialized medical fields, such as hematology, has been highlighted by the initiatives of the ASH Research Collaborative for Multiple Myeloma and Sickle Cell Disease [6]. This highlights the hematology community's increasing interest in and need for standardized, interoperable data formats. Acknowledging this need, our recent publication introduced a synthetic dataset offering realistic clinical trial data for Acute Myeloid Leukemia (AML) [7]. While this dataset represents a valuable resource, its integration into healthcare systems and research platforms could be expanded via standardized data formats. This paper presents the transformation of the AML trial data into OMOP and FHIR formats and makes it publicly available.

2. Methods

In our recent work, we leveraged two state-of-the-art generative AI models, CTAB-GAN+ [8] and Normalizing Flow (NFlow) [9], to create synthetic datasets representing 1606 AML patients from data collected across four multicenter clinical trials [7]. These models captured the complexity of the original datasets, including key demographic, laboratory, molecular, and cytogenetic variables, and patient outcomes, ensuring high utility and fidelity. The generated data comprises 87 variables per patient, including demographic information (SEX and AGE), diagnostic indicators (AMLSTAT and EXAML), outcomes (OSSTAT, EFSSTAT, OSTM, EFSTM, CR1), laboratory values (HB, WBC, PLT), 50 binary molecular genetic variables, and 24 binary cytogenetics variables. The datasets exhibit realistic variability, including missing data.

Building on the foundational work by Ahmadi et al. [10], which provided a mapping of original AML data to OMOP, we applied their existing mapping table and the Extract-Transform-Load (ETL) process to convert our synthetic datasets to the OMOP format. This conversion was executed using R programming language. The mapping utilized LOINC, UCUM, SNOMED CT terminologies, and OMOP Genomic vocabulary [11] for accurate data representation. The OMOP tables are provided at <https://zenodo.org/records/10913060>, while the mapping table and ETL code are accessible at <https://github.com/NajiaAhmadi/AML-Synthetic-data-OMOP-version>.

For the FHIR mapping, we developed a new mapping table based on the OMOP conversion, using only standard FHIR resources. One author with a Medical Informatics background conducted the initial mappings by searching the SNOMED and LOINC databases for relevant codes. When specific codes were unavailable, broader concepts were used, with detailed information provided in the text fields. For ambiguous matches, less specific codes were selected. A second author reviewed the initial coding, and a hematologist was consulted in cases of uncertainty. We used three standard FHIR resources: *Patient* for demographics, *Condition* for disease/remission, and *Observation*

for laboratory, molecular, cytogenetic, and outcome variables. As we did not find any direct way to encode the subtype of AML (de novo, sAML and t-AML), we added this information in the text field of the AML condition resource. *Condition* resources were created when the condition was present; otherwise, no resource was created. For a single patient up to three conditions were created: 1) AML disease with the subtype information, 2) Extramedullary Acute Myelogenous Leukemia (EXAML), and 3) First complete remission (CR1). We noted the absence of 'effective time' and 'performer' details in our *Observations*, as this information was not available. This limitation led to warnings in the validation process but did not compromise the dataset's integrity.

A challenge encountered during mapping was the presence of age in our dataset, which cannot be directly aligned with OMOP and FHIR standards. To address this, we generated hypothetical birthdates by subtracting each patient's age from a hypothetical fixed study end date (January 1, 2010), ensuring compliance with OMOP and FHIR standards. This workaround was documented in the human-readable text fields within FHIR to maintain transparency. In our FHIR data, we compiled all resources related to a single patient into a Bundle, enhancing accessibility and interoperability. The FHIR datasets are published on <https://zenodo.org/records/10912936>, and the mapping table and code are available at <https://github.com/waldemar93/synthetic-aml-data-to-FHIR>.

3. Results

OMOP datasets: The transformed data in the OMOP format are only briefly represented in the following because the main mapping procedure was already conducted in detail in Ahmadi et al. [10] and we here solely provide the synthetic version of the dataset. The ETL process was utilized to convert our generated hematology datasets into the OMOP format, encompassing data from 1,606 patients. The integration of the synthetic data into the OMOP format and the utilized OMOP concept ids generated four tables, namely *Patient*, *Condition*, *Observation*, and *Measurement* for each data generation method.

FHIR datasets: This section details our approach to mapping the synthetic datasets to the FHIR format. Here, we were able to map 30 out of 50 molecular genetic variables directly to their respective LOINC codes. For four additional variables an almost exact mapping was possible; *IDH1* and *IDH2*, as well as *FLT3I* and *FLT3T*, were mapped to a LOINC code representing an analysis of both variables at the same time. We used the "text" field to specify the variable of interest. The remaining 16 variables were mapped to a generic LOINC code, representing a generic gene testing panel. Again, the "text" field was used to provide information about the specific variable that was tested. Cytogenetic variables indicating Monosomies (5), Trisomies (2) or the deletion of a chromosomal part (4) were mapped to SNOMED CT, while translocations (9) and inversions (2) were mapped to LOINC. For four out of five monosomies in our dataset, there was no specific SNOMED CT code, so we used a general monosomy code and specified the monosomy in the "text" field, respectively. For two translocations and one inversion, we did not find a fitting code, so we used the generic "Chromosome analysis panel by FISH" code instead. The binary source value for each genetic observation (cytogenetic and molecular) was mapped through "valueCodableConcept" to *Detected* or *Not Detected* (SNOMED CT) with a standard "interpretation" code of *Positive* or *Negative*. In case the value was missing, we used *Unknown* as the reason for the absence of the data. In the original synthetic datasets, we had two variables for the karyotype: one

denoting a complex karyotype and one denoting a normal karyotype. We combined these two variables and mapped them to one LOINC code, representing a karyotype observation. We used the standard “interpretation” code *Abnormal* to represent complex and *Normal* to represent normal karyotype. For a lack of better fit, we used the *Observation* resource for the outcome variables Overall survival (time and status) and Event-free survival (time and status). For the two variables representing time, we used the SNOMED CT code *Survival time* using the text field to specify the variable. For the status variables, we did not find any closely fitting codes and used a broad *General clinical state* SNOMED CT code instead, specifying the variables again in the “text” field. We did not find any reasonable interpretation codes from the standard FHIR codes. Therefore, we used a non-standard SNOMED CT codes *Alive* or *Dead* for the Overall survival status. We added through the “text” field that *Alive* might also mean censored. For the Event-free survival status, we used just a textual interpretation. These four outcome observations were considered of type *Survey*, while all others were considered *Laboratory*. For all non-binary variables, we used UCUM for the unit representations.

The developed FHIR resources were validated using the online FHIR validator (<https://validator.fhir.org/>), resulting in no errors, which indicates adherence to the used FHIR standards. However, since we utilized two non-standard SNOMED Interpretation Codes (*Alive* and *Dead*) for the overall survival status, this prompted warnings from the validator. Following the best practices for FHIR development, a human-readable text in the form of HTML was added to each FHIR resource. We organized all resources regarding a single patient in a FHIR bundle, thus facilitating a cohesive, accessible, and interpretable format within the FHIR framework. We used the Bundle type *Transaction*, so that the resources can be uploaded with ease to a FHIR server.

4. Discussion

The provided OMOP and FHIR resources described in our manuscript offer added value to developers and clinicians within the field of hematology. For clinicians, the immediate benefit lies in the availability of interoperable and realistic datasets that are able to closely mimic a clinical trial scenario. These synthetic datasets can be combined with local data in the OMOP format to foster comparative observational multi-center research. For developers, the FHIR format provides a valuable source for developing complex Health Informatics applications, e.g., via the SMART-on-FHIR specification [12]. The public availability of these datasets in both OMOP and FHIR formats, coupled with accessible mapping tables, creates a versatile resource. We are convinced that the provided datasets can facilitate the development of more nuanced and precise analytical tools and algorithms, enhancing predictive modeling, data visualization, and machine learning applications in hematology, especially in AML.

In this work, we chose FHIR standard resources over customized profiles to ensure broad compatibility and interoperability across global healthcare systems. This approach allows for easier data sharing and integration among healthcare providers, researchers, and developers, irrespective of the technology or platform used. Customized FHIR profiles, while potentially more tailored to specific needs, require additional development iterations, as reported in related works [13]. Our results indicate that not all variables can be directly mapped using standard LOINC and SNOMED CT terminologies, leading to the use of broader concepts with textual explanations, which might impact resource usability. Additionally, we noted a limitation in how laboratory,

genetic findings, and outcome variables were represented as generic observation resources. To overcome these challenges, we plan to reassess our FHIR to align closer with the mCODE standard [5], allowing the use of HGNC IDs for genetic observations, facilitating a more streamlined and accurate mapping procedure for genetic data.

5. Conclusions

In summary, we demonstrate the transformation of synthetic AML patient datasets into OMOP and FHIR formats, enhancing the interoperability and accessibility of hematology research data. By providing these standardized datasets publicly, this work paves the way for advancements in health informatics applications, offering a valuable resource for both clinical and research endeavors in the field of hematology.

References

- [1] Kokosi T, Harron K. Synthetic data in medical research. *BMJ Med.* 2022 Sep 26;1(1):e000167. doi: 10.1136/bmjmed-2022-000167.
- [2] Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. *PLOS Digit Health.* 2023 Jan 6;2(1):e0000082. doi: 10.1371/journal.pdig.0000082.
- [3] Fennelly O, Moroney D, Doyle M, Eustace-Cook J, Hughes M. Key interoperability Factors for patient portals and Electronic health Records: A scoping review. *Int J Med Inform.* 2024 Mar;183:105335. doi: 10.1016/j.ijmedinf.2023.105335.
- [4] Bennett AM, Ulrich H, van Damme P, Wiedekopf J, Johnson AEW. MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format. *J Am Med Inform Assoc.* 2023 Apr;30(4):718-725. doi: 10.1093/jamia/ocad002.
- [5] Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. *JCO Clin Cancer Inform.* 2020 Oct;4:993-1001. doi: 10.1200/CCI.20.00059.
- [6] Wood WA, Marks P, Plovnick RM, Hewitt K, Neuberger DS, Walters S, Dolan BK, Tucker EA, Abrams CS, Thompson AA, Anderson KC, Kluetz P, Farrell A, Rivera D, Gertzog M, Pappas G. ASH Research Collaborative: a real-world data infrastructure to support real-world evidence development and learning healthcare systems in hematology. *Blood Adv.* 2021 Dec 14;5(23):5429-5438. doi: 10.1182/bloodadvances.2021005902.
- [7] Eckardt JN, Hahn W, Röllig C, Stasik S, Platzbecker U, Müller-Tidow C, Serve H, Baldus CD, Schliemann C, Schäfer-Eckart K, Hanoun M, Kaufmann M, Burchert A, Thiede C, Schetelig J, Sedlmayr M, Bornhäuser M, Wolfen M, Middeke JM. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *NPJ Digit Med.* 2024 Mar 20;7(1):76. doi: 10.1038/s41746-024-01076-x.
- [8] Zhao Z, Kunar A, Birke R, Van der Scheer H, Chen LY. CTAB-GAN+: enhancing tabular data synthesis. *Front Big Data.* 2024 Jan 8;6:1296508. doi: 10.3389/fdata.2023.1296508.
- [9] Papamakarios G, Nalisnick E, Rezende DJ, Mohamed S, Lakshminarayanan B. Normalizing flows for probabilistic modeling and inference. *J Mach Learn Res.* 2021;22(57):1-64.
- [10] Ahmadi N, Zoch M, Guengoze O, et al. How to customize Common Data Models for rare diseases: an OMOP-based implementation and lessons learned. 2023 Dec 08. Preprint (Version 1) available from: Research Square.
- [11] Kaduk D, Komar V, Golozar A, Robinson P, Wagner AH, Gurley M, et al. Genomic data harmonization through the OMOP standardized vocabularies. In: *Proceedings of the 2020 OHDSI Global Symposium*; 2020 Oct; Virtual Symposium. p. 18-21.
- [12] Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Inform Assoc.* 2016 Sep;23(5):899-908. doi: 10.1093/jamia/ocv189.
- [13] Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: Authoring FHIR implementation guides based on WHO recommendations. *J Biomed Inform.* 2021 Oct;122:103891. doi: 10.1016/j.jbi.2021.103891.