UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Zarrow: a modular framework for clinical cohort construction and machine learning benchmarking on multimodal medical datasets [Abstract]

Pierre Aublin, Stephanie Kühne, Dario Bongiovanni, Sebastian Zaunseder

Angaben zur Veröffentlichung / Publication details:

Aublin, Pierre, Stephanie Kühne, Dario Bongiovanni, and Sebastian Zaunseder. 2026. "Zarrow: a modular framework for clinical cohort construction and machine learning benchmarking on multimodal medical datasets [Abstract]." *Biomedical Engineering / Biomedizinische Technik* 71 (s1). https://doi.org/10.1515/bmt-2026-1001.

Nutzungsbedingungen / Terms of use: **CC BY 4.0**

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Zarrow: A modular Framework for Clinical Cohort Construction and Machine Learning Benchmarking on Multimodal Medical Datasets

Pierre Aublin, Chair for Diagnostic Sensing, University of Augsburg, Augsburg, Germany, pierre.aublin@uni-a.de
Stephanie Kühne, Department of 1st Internal Medicine, University Hospital Augsburg, Augsburg, Germany, stephanie.kuehne@med.uni-augsburg.de
Dario Bongiovanni, Department of 1st Internal Medicine, University Hospital Augsburg, Germany, dario.bongiovanni@med.uni-augsburg.de
Sebastian Zaunseder, Chair for Diagnostic Sensing, University of Augsburg, Augsburg, Germany, sebastian.zaunseder@uni-a.de

Introduction

AI model development in clinical research is hindered by a lack of reproducibility, heterogeneous data formats, and a lack of pipelines to construct standardized cohorts. While public datasets like MIMIC provide a valuable data resource, the conversion of the raw relational tables into analysis ready datasets results in error-prone and time-consuming pipeline redevelopments. The extension of datasets to integrate new modalities like waveforms, ECG, X-ray, or text notes further increases the complexity of this process.

Methods

We present a modular Python framework for standardized reproducible clinical cohort construction and machine learning dataset building. The framework allows for flexible cohort definition through composable filtering operations. It supports scalable processing and downloading through chunk data loading, and efficient storage through Zarr format for large scale datasets. The framework is designed to process multiple data modalities including static tabular data and time-series. For multimodal data integration, the framework avoids unnecessary full dataset re-downloads by allowing incremental data access and reusing the previously constructed clinical cohorts.

Results

We demonstrate the framework's use on MIMIC-IV by building a cohort of patients with acute myocardial infarction based on ICD codes. From this cohort, a downstream dataset is derived, integrating demographic variables, laboratory measurements, and defining 28-day mortality endpoint based on admission and death timestamps. The resulting pipeline fully automates the process getting from raw data to a dataset which can be plugged into machine learning and deep learning frameworks.

Conclusion

This work provides an extensible framework to standardize clinical data processing across multiple datasets for machine learning workflows. By reducing the implementation overhead, the framework eases faster and more reliable implementation of machine learning models in healthcare. Our first demonstration focused on demographic variables and laboratory measurements. Future works direct at multimodal, i.e. signals' and images', integration and representation learning approaches.

Open Access. © 2026 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.