Contribution ID: **152** Type: **Presentation**

A General-Purpose Framework for Structured, Reproducible, and Transparent Data Harmonization

Wednesday 15 October 2025 14:33 (11 minutes)

Despite efforts to improve the availability and accessibility of research datasets, interoperability remains a serious barrier to reuse. Data harmonization, the process of aligning data from disparate sources to a standardized schema, plays a key role in addressing data heterogeneity and enabling integration and reuse. Consider a data scientist curating patient blood-glucose measurements as a precursor to secondary data analysis. Secondary data analysis often requires the integration of multiple datasets into an aggregate dataset large enough to provide sufficient statistical power to investigate a hypothesis or to provide adequate training data for a computational model. If the datasets are heterogeneous, the data scientist must first align them by implementing a data harmonization strategy.

Data harmonization entails mapping the schema of an original dataset to a standardized target schema, often chosen by the consumer of the data or established by a data repository. For example, the data scientist may need to compile a dataset that represents blood-glucose measurements using a categorical variable for clinically-relevant blood-glucose levels (e.g., hypoglycemic, normal, prediabetic, and diabetic). If the data scientist encounters a dataset that instead records numerical blood-glucose levels (e.g., in units of mg/dL), those data must be mapped from the numerical to the categorical schema to enable integration.

Recent developments in the literature of data harmonization have focused on establishing best practices for implementing data harmonization [1,2]. However, current harmonization practices suffer from limitations in reproducibility and flexibility, and existing tools to support harmonization are either specialized to domain-specific datasets or part of large integrated data management systems. Modern harmonization methods rely on hard-coded, manually maintained scripts, which limit adaptability when data standards and research objectives evolve. Consequently, the revision of outdated data representations or harmonization protocols, especially those developed early in a project, can be prohibitively labor-intensive. Moreover, harmonization protocols are conventionally documented in text, which leaves room for interpretation when implemented by a reader, resulting in harmonized datasets with opaque provenance and limited reproducibility.

To address these limitations, we have developed a general-purpose data harmonization framework that emphasizes reproducibility and transparency [3]. Our framework achieves reproducibility using a novel strategy of building data transformations from standardized building blocks called "primitive"operations. For example, to transform numerical blood-glucose measurements to labeled categories, the data scientist would leverage the "Bin"primitive, which assigns numerical values to categorical labels using histogram ranges, e.g., hypoglycemic (<70 mg/dL), normal (70–99 mg/dL), prediabetic (100–124 mg/dL), and diabetic (≥125 mg/dL). Our framework defines a standard vocabulary of named primitive operations, including, among others, the "Bin" primitive above, the "ConvertUnits"primitive for performing unit conversions (e.g., from mg/dL to mmol/L), and the "Round"primitive for rounding a decimal value to a specific precision (e.g., from 1.15 to 1.2 at one significant digit of precision). Each primitive operation can be parameterized to fine-tune its behavior, such as by providing labels and their corresponding numerical ranges for the "Bin"primitive as demonstrated by the harmonization of blood-glucose levels. Moreover, primitives can be composed to achieve complex transformations, such as by performing unit conversion followed by decimal rounding to conform to a standardized precision level. Our declarative language of primitive operations provides standardization for representing previously ad hoc data transformations, while the abilities to parameterize and compose primitives provide the flexibility to build complex harmonization procedures.

The standardization established by the vocabulary of primitives affords an additional advantage: the ability to serialize harmonization protocols in a machine-readable format. A data transformation implemented using our framework is completely specified by naming the primitives that compose the transformation and by

including the parameters used by each primitive. A harmonization protocol implemented as part of a computational harmonization workflow can be stored externally (e.g., in JSON format) and reconstructed from its serialization, thereby enabling the exact reproduction of a previously harmonized dataset and, in turn, guaranteeing reproducibility for downstream secondary data applications.

Additionally, our framework records all executed harmonization transformations in order to provide traceability for the resulting harmonized and integrated dataset. A different researcher can obtain the integrated blood-glucose dataset compiled by the data scientist in the earlier example and inspect the provenance of an individual element within the integrated dataset to trace its origin and identify the transformations applied to it as part of harmonization. This traceability allows the researcher to determine whether the harmonized dataset suits their research goals or whether to consider the original data under an alternative harmonization strategy.

In this presentation, we will detail the conceptual models and technical components of our harmonization framework and showcase real-world results from applying the framework within the RADx Data Hub, a repository established by the National Institutes of Health (NIH) as part of the Rapid Acceleration of Diagnostics (RADx) program for hosting research data collected during the pandemic [4]. The diversity of research programs, methods, and data types represented in the RADx repository demonstrates the framework's effectiveness in achieving principled, reproducible, and transparent data harmonization for complex heterogeneous data.

References

1. Fortier, I. et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. Int. J. Epidemiol. 46, 103–105 (2017).

2. Cheng, C. et al. A general primer for data harmonization. Sci. Data 11, 152 (2024).

3. Yu, J. K., et al. A general-purpose data harmonization framework: supporting reproducible and scalable data integration in the RADx Data Hub. Preprint at https://doi.org/10.48550/arXiv.2503.02115 (2025).

4. Martinez-Romero, M. et al. RADx Data Hub: a cloud platform for FAIR, harmonized COVID-19 data. Preprint at https://doi.org/10.48550/arXiv.2502.00265 (2025).

Primary author: YU, Jimmy (Stanford University)

Co-authors: Dr MARTÍNEZ-ROMERO, Marcos (Stanford University); Prof. MUSEN, Mark (Stanford University); Dr HORRIDGE, Matthew (Stanford University); Dr AKDOGAN, Mete (Stanford University)

Presenter: Prof. MUSEN, Mark (Stanford University)

Session Classification: Presentations Session 7: Open research through Interconnected, Interoperable, and Interdisciplinary Data

Track Classification: SciDataCon2025 Specific Themes: Open research through Interconnected, Interoperable, and Interdisciplinary Data