



Data that Defends Itself: Machine Learning for Error-Proof ADaM Submissions

Bhargav Koduru

PHUSE SDE BOSTON 2025

11 June 2025

Introductions



Bhargav Koduru

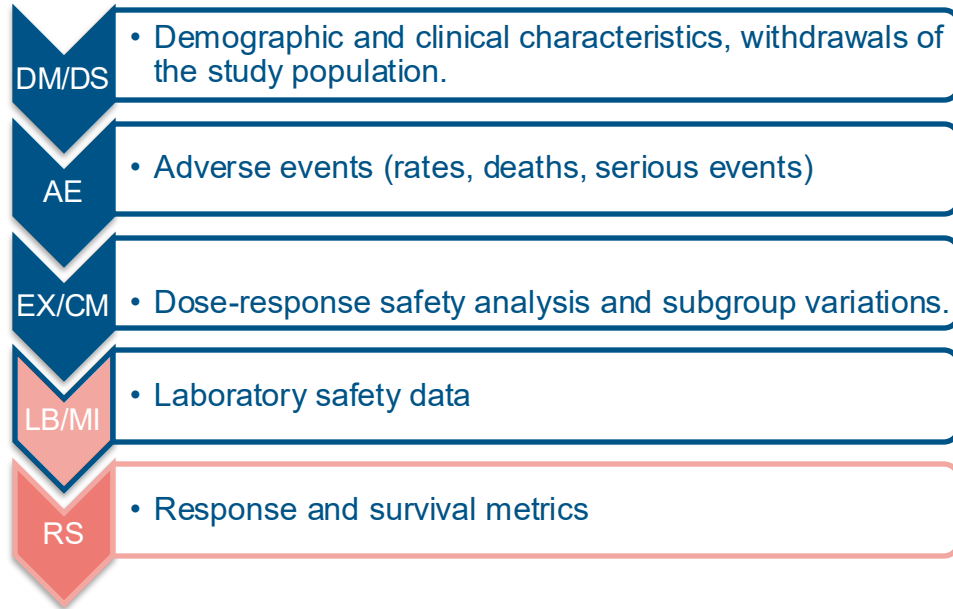
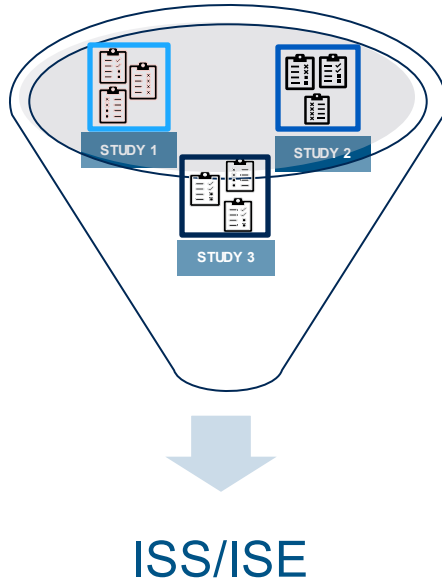
Senior Manager, Statistical Programming, Clinical Data Standards

- Driving Clinical and Operational Data Standards at Medidata.
- Over a decade of experience at Seattle Genetics (now Pfizer), Sarepta, and Bioclinica (now Clario).
- MS Quantitative and Systems Biology, UCM
- MSc (Medical) Biochemistry, Manipal University
- Email: bhargavkoduru@gmail.com

ISS/ISE: A Brief Overview

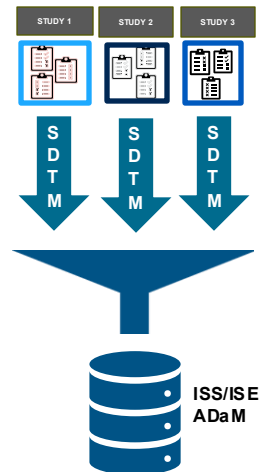
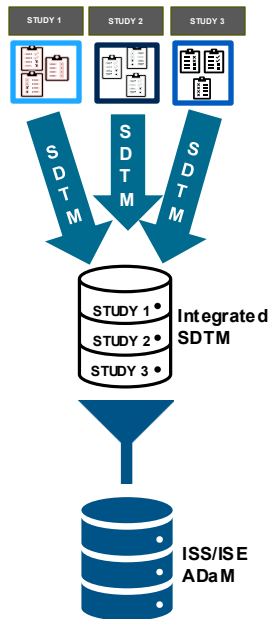
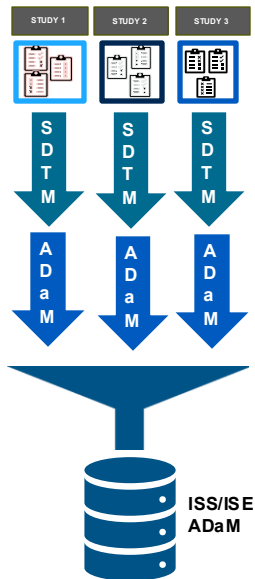
ISS/ ISE

Data from all relevant clinical trials are pooled, integrated and results compared across studies.



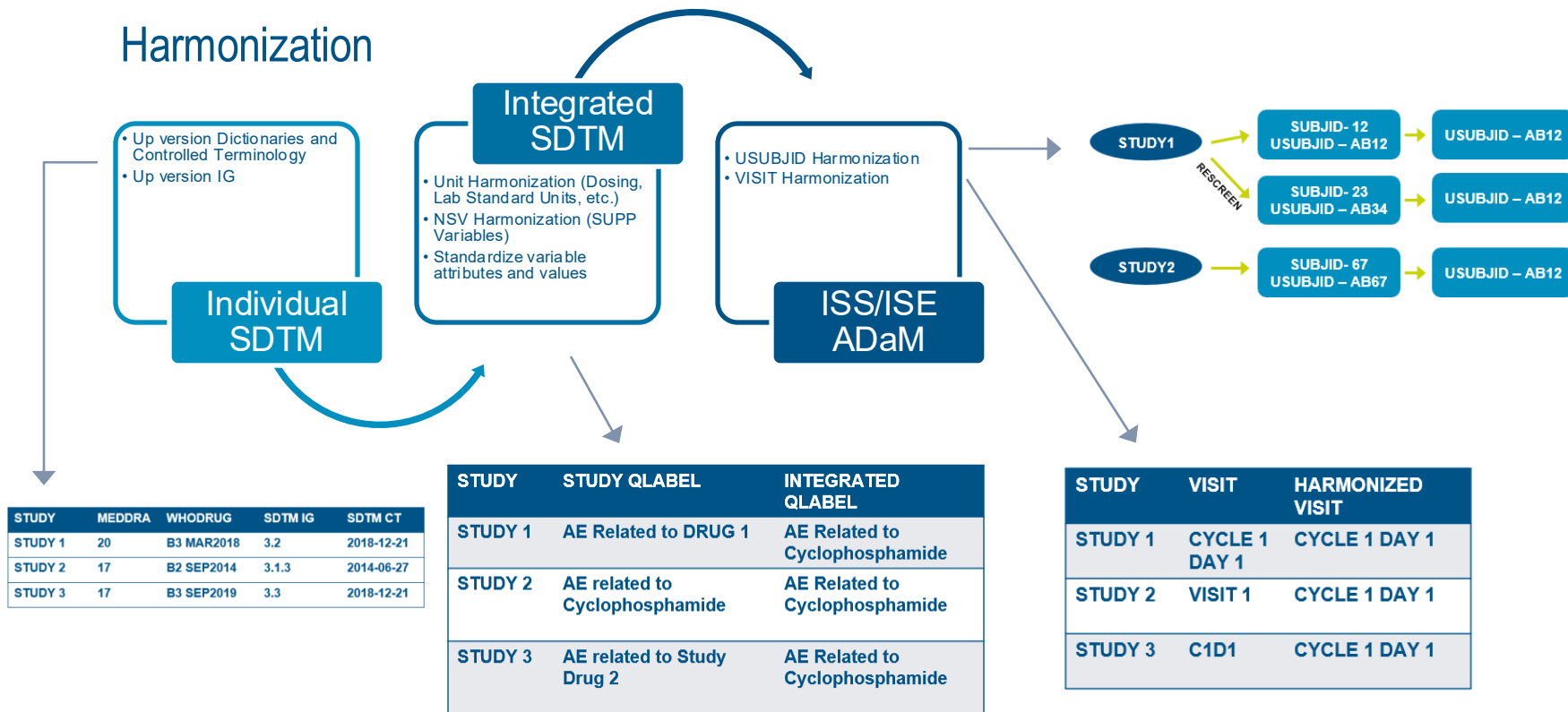
ISS/ISE

Integration



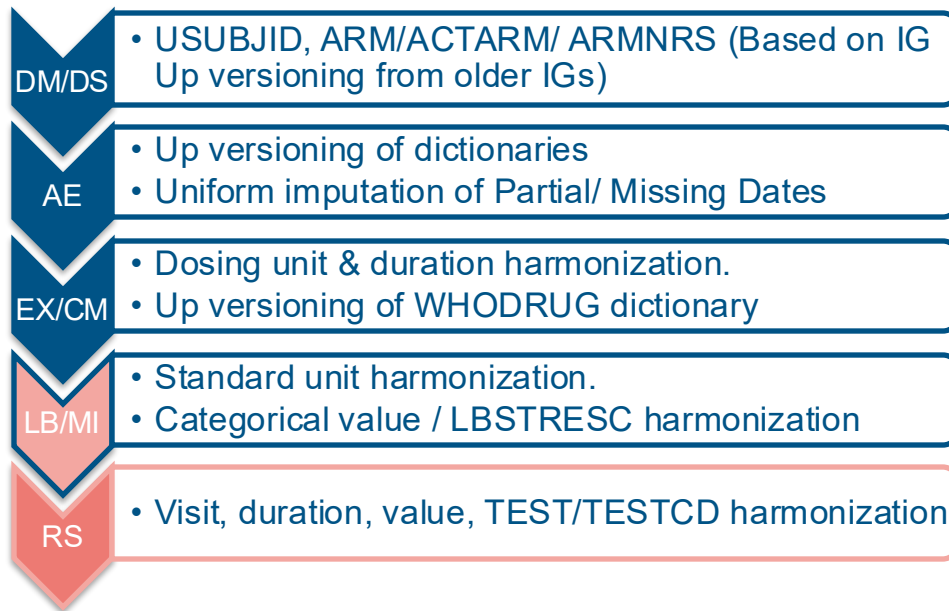
ISS/ISE

Harmonization



ISS/ISE

Quality control of data



Machine Learning / AI to the Rescue.

Outliers

Types of detection

Outliers are extreme values that depend on context, model, and data type.



Univariate Outliers

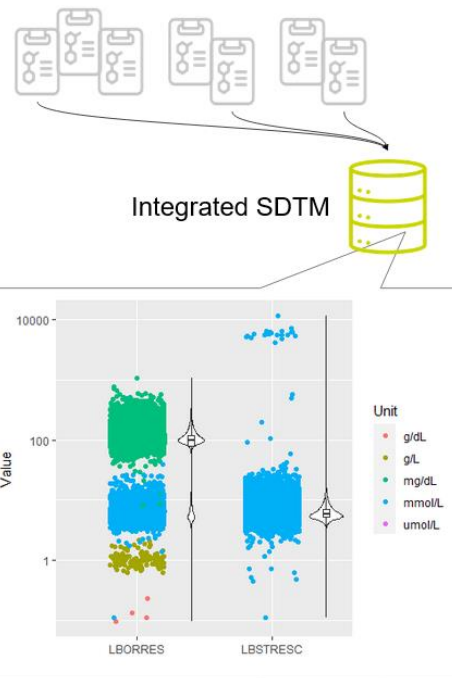
- Extreme values within a single population or variable

Bivariate Outliers

- Extreme values conditional on other variables (e.g., across studies or patient characteristics)

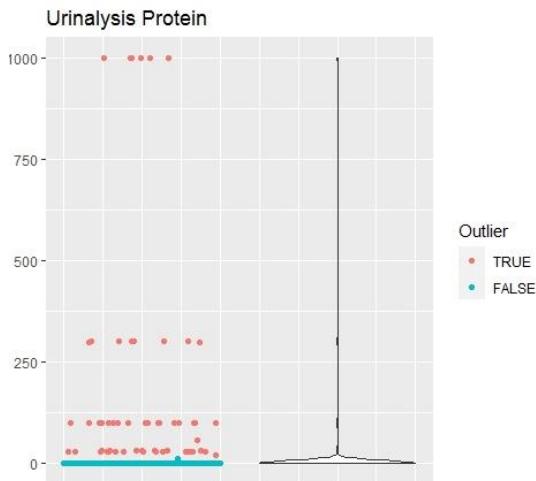
Subject-Specific Time-Series Outliers

- Values showing unusual temporal changes for an individual, compared to other patients



Univariate Outliers

Overlay of Univariate with standard conversion factors



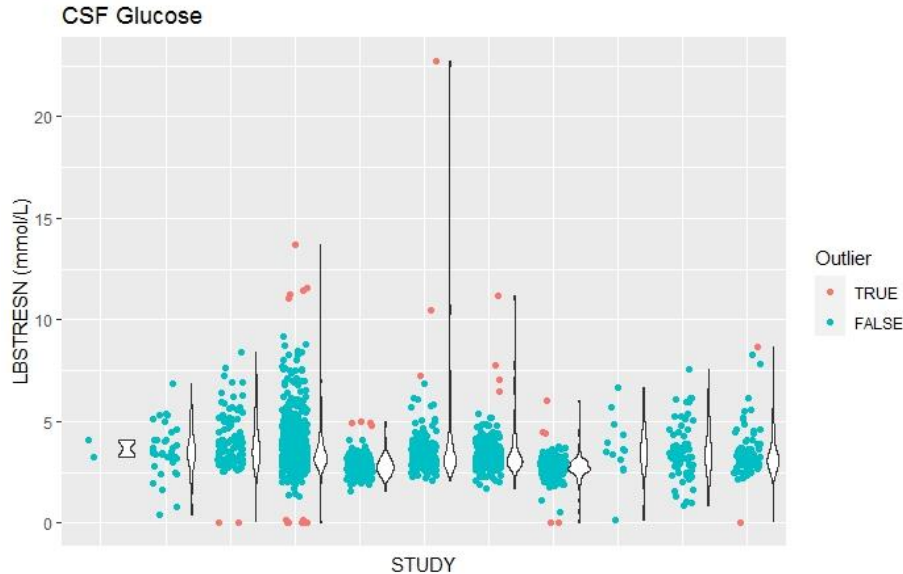
LBTEST	LBORRES	LBORRESU	LBSTRESN	LBSTRESU	Derived Conversion Factor	True Conversion Factor	Expected LBSTRESN
Protein	100	mg/dL	1000	g/L	10	0.01	1

A few number of data points (~0.8%) are found to have abnormally high levels of total protein in urine samples, where majority data points are close to 0 g/L.

For context, Normal daily urine protein excretion is typically less than 150 mg/day (or <0.15 g/day)

Bivariate Outliers

Dynamic adjustment based on correlation of variables

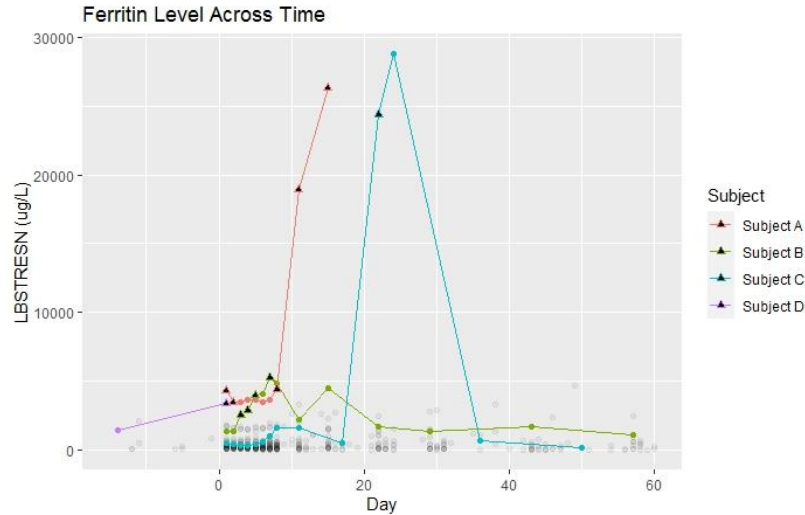


One study has a higher amount of variability, and it has a higher threshold in flagging outliers (>10 mmol/L).

Around 1.7% of data points were flagged as outliers, half the number of data points flagged when all studies were pooled together.

Indicates that study-specific thresholds and variability can significantly impact outlier identification

Time series outliers

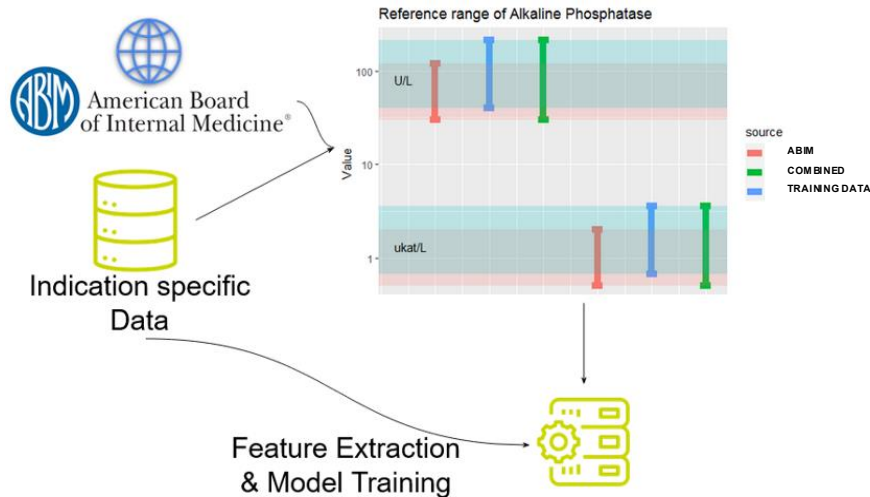


All patients have relatively stable Ferritin level throughout the trials, except the three patients as shown in the graph.

Two patients have abnormally high value as time progressed, with sudden spikes.

Kho, S.J., Strait, C., **Koduru, B.**, & Burke, K. (2024). Systems and Methods for Maintaining Data Integrity in a Health Analysis Platform by Assessing and Modifying Time-Series Outliers in Filtered Healthcare Data. (U.S. Provisional Application, Pending, Application No. 18732347). U.S. Patent and Trademark Office.

Knowledge Extraction Framework



A sensible reference range derived from Indication specific training data and ABIM, along with unit usage frequency information, are used to train a classifier to detect mislabeled unit.

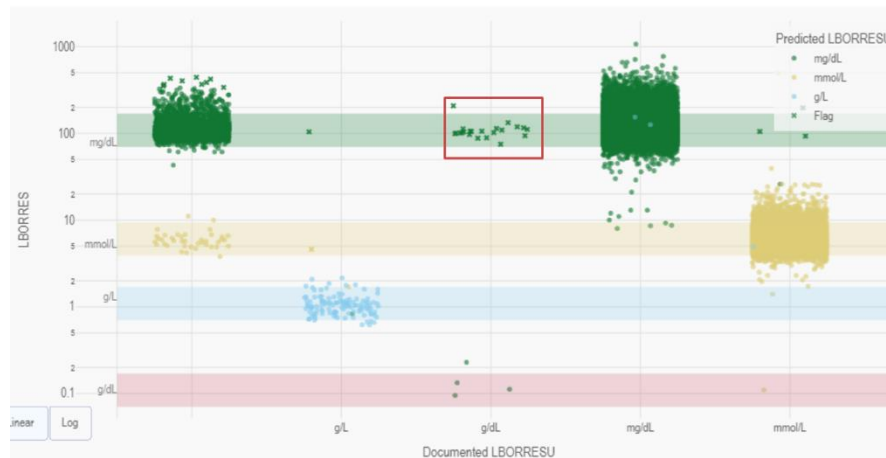
Kho, S J., Strait, C., **Koduru, B.**, & Burke, K. (2024). Systems and Methods for Maintaining Data Integrity in a Health Analysis Platform by Assessing and Modifying Physiological Measurements Based on Filtered Healthcare Data. (U.S. Provisional Application, Pending, Application No.18732403). U.S. Patent and Trademark Office.

Knowledge Extraction Framework

Incorrect/Missing Unit Detection

Data points labeled as “g/dL” predicted to be “mg/dL”

Correct units were predicted for data points with missing units.



References

- Kho, S., & **Koduru B.** (2024, October 23-24). *Multi-trial Laboratory Data Quality Assessment and Profiling of SDTM Data through a Hybrid Statistical-Machine Learning Approach* [Conference presentation]. CDISC + TMF US Interchange 2024, Phoenix/Scottsdale, AZ, United States
- Kho, S., & **Koduru B.** (2024, June 3-5). [Unlocking the Potential of Historical Clinical Trials: A Hybrid Statistical and Machine Learning Approach for Laboratory Data Quality Assessment](#) [Conference presentation]. PHUSE/FDA CSS 2024, Silver Spring, MD, United States
- CP, J., & Vijayan, A., Overcoming Challenges of your Integrated Summary ISS/ISE Submissions, PHUSE 2022.
- Integration Strategies in Support of ISS/ISE Submissions, PHUSE White Paper