Anomaly detection should depend on **context: outliers** differ by **age and gender**

Title: Detecting implausible measurement values in observational data using machine-learning based outlier detection methods

Background: Standardizing real-world data to the OMOP Common Data Model (CDM) enables systematic evaluation of data quality, with the OHDSI DataQualityDashboard (DQD) providing rule-based checks aligned to the Kahn et al. framework for conformance, completeness, and plausibility. However, these predefined checks may miss unexpected or context-specific anomalies in measurement data. In this study, we evaluate concept-level plausibility using both univariate and machine learning approaches using systolic blood pressure (SBP) as a representative clinical example. Univariate methods included Tukey fences (IQR-based), z-score, modified z-score (MAD-based), percentile-based thresholds, and robust standard deviation rules to identify extreme values across global distributions. In parallel, we applied two unsupervised machine learning methods: Isolation Forest (IF), which isolates anomalies using an ensemble of random trees such that anomalous observations require fewer splits to separate, and Gaussian Anomaly Detection (GAD), which models the data distribution and flags observations with low probability under the assumed distribution. Together, these approaches enable comparison of rule-based and data-driven methods for identifying implausible measurement values and augmenting existing rule-based data quality assessments.

Univariate methods converge on a consistent range for plausible systolic blood pressure values

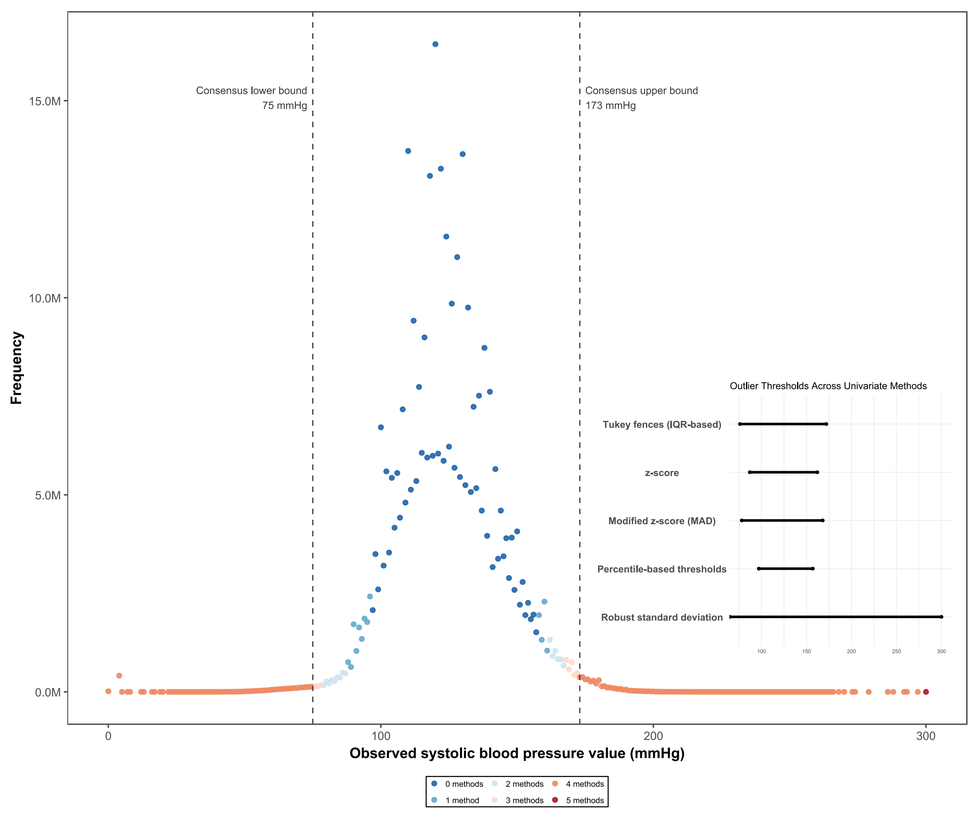


Figure 1: Distribution of systolic blood pressure values without outliers identified by five univariate methods. Values flagged by >= 3 methods define lower and upper bound for plausible measurements.

Outlier detection methods agree overall, but become more sensitive when stratified by age and gender

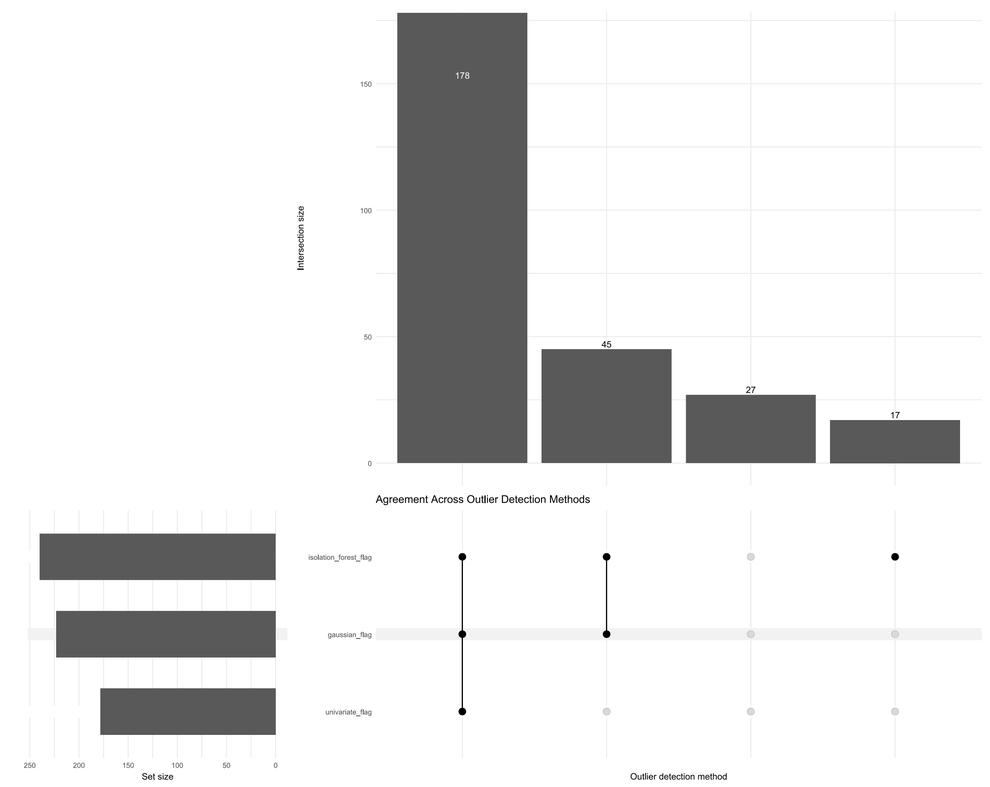
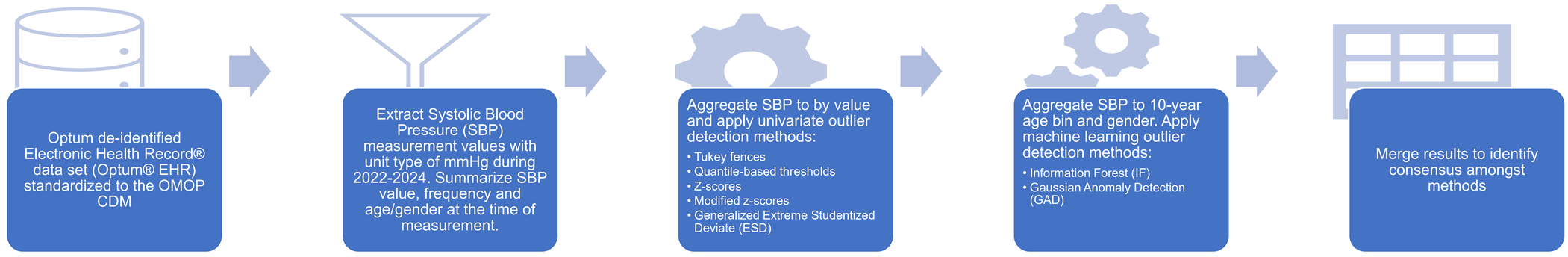


Figure 2: UpSet plot showing overlap of systolic blood pressure (SBP) values flagged as outliers by global univariate methods and stratified machine learning approaches (Isolation Forest and Gaussian Anomaly Detection). Set size bars indicate the total number of SBP values flagged by each method, while intersection bars represent the number of values identified by specific combinations of methods..

Methods



Future Work will explore how AI-driven methods can be used to assess and improve data quality at scale across all measurement types and, ultimately, across all domains. We invite discussion on how these approaches can be integrated into the federated OHDSI networks.

Anthony G. Sena[1,2], Peter R. Rijnbeek[1], Aniek F. Markus[1]

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Johnson & Johnson, Raritan, NJ, USA