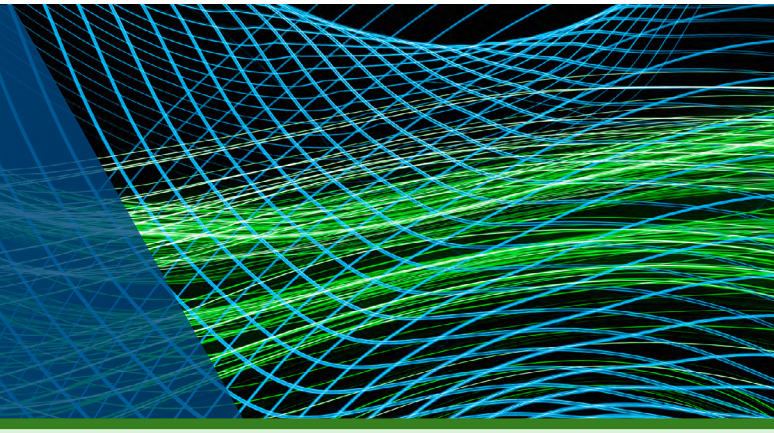
Data Science @ Westat

Data Integration and Harmonization



Westat achieves effective and efficient data integration and harmonization of data from disparate sources, processes data of varying file formats, naming conventions, and columns, and transforms them into a cohesive data set. These data are then available for streamlined analyses based on a common standard.

Wide-Ranging Expertise

Data sources may be similar data from different sources (e.g., clinical data); complementary data used to create single measures or results; or the same types of data from different studies.

- In clinical and epidemiological research studies, we integrate and harmonize data from electronic health records (EHR) with prospectively collected clinical data to support novel analyses.
- In public health research, we support disease surveillance efforts through the linkage of clinical data with administrative data sources, such as the National Death Index (NDI), state cancer registry data and the American Community Survey (ACS) or US Census data.
- We employ a range of common data models, including the Observational Medical Outcomes Partnership (OMOP) model, to facilitate harmonization of data from disparate sources.
- For clinical trials, we employ CDISC standards (CDASH, STDM, and ADaM) to streamline and automate clinical trial data processing, enhance quality, and support regulatory submissions.
- We develop complex extract-transform-load (ETL) pipelines to facilitate automated mapping between differing data formats to define and apply data harmonization rules and to pool participant-level data from multiple sources. These pipelines adhere to data security standards and privacy/confidentiality standards.

Innovative and Efficient Approaches

Westat has developed pipelines to support data ingestion from different sources as well as standardized algorithms to conduct integration of large volumes of data for dataset creation.

- Adoption of common data models and common data elements, including OMOP, NIH Common Data Elements, PhenX measures for data collection, and PROMIS measures.
- Application of EHR-related standards and terminology systems to facilitate cross-system analysis (e.g., FHIR/HL7, ICD-9/10-CM, CPT, SNOMED-CT).
- Development of systems to integrate imaging data with clinical data using industry standards (e.g., DICOM).
- Metadata-driven approaches to data harmonization, including use of configurable rules-based ETL pipelines and searchable metadata repositories.
- Application of multiple data harmonization approaches, including algorithmic data transformation (e.g., mapping of categorical variables with different but compatible categories to a common standard), mapping of continuous variables to common units, scale standardization, construction of derived measures, and imputation of missing data.

Illustrative Projects

The Recipient Epidemiology and Donor Evaluation-IV Pediatric (REDS-IV-P) Research Program. In a project for the National Heart, Lung, and Blood Institute (NHLBI) designed to respond to potential threats to the safety of the blood supply and to address emerging research needs in transfusion medicine, Westat built a comprehensive Research Data Warehouse (RDW) comprised of electronic health record patient data from 22 hospitals, blood donor and blood product data from 4 blood centers, and omics data on donor samples using a RED-IV-P customized genotype array. Data are sent to Westat using a secure data pipeline created by Westat, are stored on the Amazon Web Services (AWS) cloud, and harmonized using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Westat analyzes large data files, processed on a high-performing Linux server, to address key research questions in blood banking and transfusion medicine and inform blood policy decisions.

For the **National Hospital Care Survey (NHCS)**, a contract funded by the Centers for Disease Control and Prevention, Westat collects patient-level data across all inpatient encounters from more than 600 U.S. hospitals. To support the processing of the EHR data, Westat data scientists developed a highly scalable system to validate and extract data from EHRs as HL7 continuity of care documents. They perform linkages of patient encounters to data from the NDI and the Centers for Medicare & Medicaid Services. Westat statisticians developed statistical algorithms to link patients across hospitals and years. We also developed a data cleaning tool using natural language processing to combine provider mentions for coding, and developed a harmonization and integration approach to combine all data from the various source types and create one database of patient encounters.

Under multiple support services contracts with the National Cancer Institute, Westat began conducting linkages to public health surveillance data in the late 1990s for over a half million cohort participants who had enrolled in the **NIH-AARP Diet and Health Study**. Westat created customized programs for state cancer registry searches to link datasets containing over 1.35 million cohort records efficiently, while controlling for specificity and sensitivity. Automated processes identified sets of highly likely matches and non-matches, resulting in a more manageable subset of matches requiring manual review. This model was used successfully for multiple rounds of linkages to 11 state cancer registries. Data was harmonized across the state cancer registries and was integrated into the existing cohort database. Westat also conducted linkages at the participant level to the National Death Index and Centers for Medicare & Medicaid Services. Linkages at the Census tract-level were conducted with the U.S. Census Bureau, American Community Survey, and air-quality monitoring data.

On the **Cancer Trials Support Unit (CTSU)** contract for NCI, Westat integrates heterogeneous databases across sites through data harmonization. Data from 540,000 participants in cancer clinical trials have been harmonized and pooled. Key to this is development of National coverage analysis templates and Electronic Medical Record (EMR) setup templates.

CONTACT

Jay Clark
Senior Statistician
240-453-2762
JayClark@westat.com