ClinEpiDB: Global Health Data Sharing, Semantic Harmonization and Exploratory Data Analysis

Cristina Aurrecoechea[1], John Brestelli[2], Brian P. Brunk[2], Danielle Callan[2], Dave Falke[1], Steve Fischer[2], Omar Harb[2], Danica Helb[2], Jay Humphrey[1], John Judkins[2], Jessica C. Kissinger[1], Nupur Kittur[1], Brianna Lindsay[2], David S. Roos[2], Sheena Shah Tomko[2], Christian J. Stoeckert, Jr[2], Steph Wever Schulman[2], Jie Zheng[2]

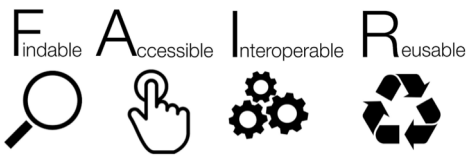
[1]*University of Georgia, Athens, GA 30602, USA,* [2]*University of Pennsylvania, Philadelphia, PA 19104, USA*

@ClinEpiDB

help@clinepidb.org

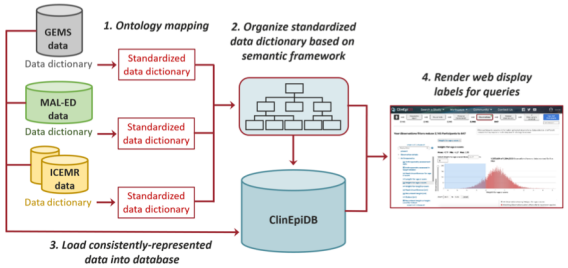
Introduction

ClinEpiDB.org is an open-access online resource that facilitates exploration and analysis of data from large-scale clinical and epidemiology studies.



ClinEpiDB promotes FAIR data through:

- **Study pages** describing study methodology and providing links to publications, case report forms, and data dictionaries
- **Semantic harmonization** which underpins queries and promotes interoperability
- **Search Wizard** facilitating data exploration
- **Search Strategies** enabling complex queries
- Tabular display of query results
- **Analysis Apps** for visualizing data
- **Data Download** for further analysis
- Tiered system controlling **data access**
- Links to corresponding data in other resources



Mapping the data variables into a robust ontological framework is critical to enable queries and analysis. Heterogeneous data variables are standardized using Open Biological and Biomedical Ontology (OBO) Foundry ontologies in a unified semantic web framework. More than 1500 different data variables are currently represented on ClinEpiDB.

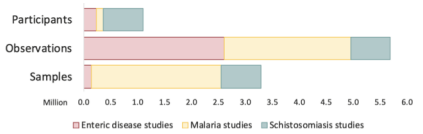
ClinEpiDB currently includes 27 datasets:

Full same-day access to 17 datasets. For datasets with access restrictions, ClinEpiDB works closely with the study team to verify requests within 1 week.

Infectious disease:
- Malaria
- Enteric diseases
- Schistosomiasis
- Respiratory diseases

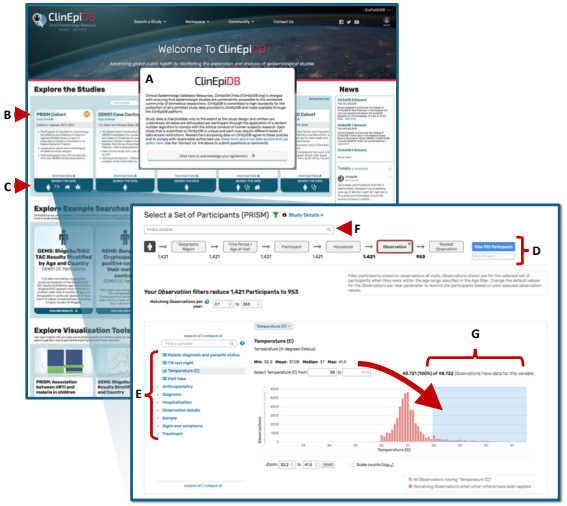
Maternal, Newborn, & Child Health

The ClinEpiDB database currently includes **> 1 million participants,** with **> 5.5 million days of observation** about associated anthropometry, demographics, and disease status and **> 3 million samples** with laboratory test results.

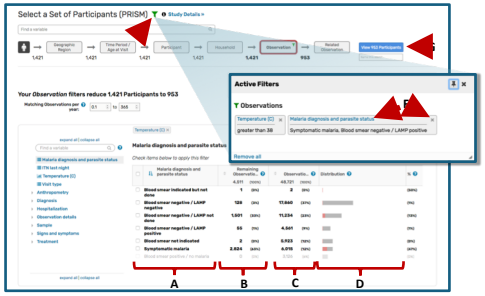


Using ClinEpiDB

Home Page: (A) Overlay appears the first time a user enters the site, providing information about appropriate use of ClinEpiDB data. **(B)** Clicking a study name opens a descriptive study page. **(C)** Clicking an icon opens the ***Search Wizard.*** **(D)** The Search Wizard categorizes the variables into discrete steps. The grey buttons let users move between steps. **(E)** The variable tree contains all variables within that step. **(F)** The "Find a variable" searches across all Search Wizard steps. **(G)** Subset data by typing or click and dragging the mouse across the range of interest.



Search Wizard: **(A)** Subset data based on categorical variables via check boxes next to the values. See how subsetting data impacts other variables by comparing the "Remaining" column **(B)** to the "Observations" column **(C)**. Data included in the subset are red on the distribution graph while the rest of the data are grey **(D)**. Use the green filter icon **(E)** to edit or remove filters **(F).** Use the blue button **(G)** to get to the ***Results Page***.



Results Page and Downloads: **(A)** Histogram icons open pop ups showing the distribution of data for that variable. Use the buttons and links to add additional variables to the table **(B)** and download the selected data **(C).** Note that studies with download restrictions will require log in and submission of a data access request **(D)**. Access the ***Analysis Apps*** for additional data visualization tools **(E).**

Analysis Apps: Vizualize results using three different Analysis Apps: Distributions, Contingency Tables (shown) or Data Summaries. **(A)** Set plot parameters based on variables of interest. **(B)** Stratify the data based on additional variables. View resulting graphs **(C)** and tables **(D)**.