Quick Start Guide for PDS-MEPS Linked Datasets

1) *What are the PDS-MEPS linked datasets, and why were they created?*

The Project Data Sphere® (PDS) platform provides the research community with broad access to de-identified patient-level clinical trial data. These data are rich in measures that characterize the clinical trials under study, their treatment protocols, and patient outcomes, and they support research topics such as the relationship between tumor growth and survival, survival prediction models based on trial design, and meta-analysis of standards of care. However, to address the confidentiality provisions inherent to the trials, data providers are required to de-identify the patient-level records by masking or removing certain demographic data. Consequently, researchers have limited ability to study the influence of health-related and socioeconomic factors, access to and use of health care services, and predisposition of health behaviors on treatment effects and patient outcomes.

The PDS-MEPS linked datasets were created to address these analytic constraints. For a selection of clinical trials on the Project Data Sphere (PDS) platform, comparator arm patients have been matched, or "linked," with similar cancer survivors from the nationally representative Medical Expenditure Panel Survey (MEPS). The result is a linked dataset containing matched pairs of PDS patients and MEPS cancer survivors. Through the linkage process, patient-level data from the clinical trials have been augmented with social, economic, and health-related variables from MEPS.

2) *What is the Medical Expenditure Panel Survey (MEPS)?*

MEPS is the United States' primary source of nationally representative, comprehensive, person-level data on health care use, insurance coverage, and expenses. MEPS has been collecting data on health care utilization and expenditures annually since 1996 and is sponsored by the Agency for Healthcare Research and Quality.

MEPS consists of a family of three interrelated surveys: Household Component (MEPS-HC), Medical Provider Component (MEPS-MPC), and Insurance Component (MEPS-IC). MEPS-IC also collects establishment-level data on insurance programs. The MEPS-HC survey is the primary source of information used in the PDS-MEPS linked datasets. Through a series of interviews with household respondents, MEPS-HC collects detailed information at the level of the individual respondent on demographic characteristics, health status, health insurance, employment, and medical care use and expenditures. These data support estimates both for individuals and for families in the United States.

The MEPS-HC consists of an overlapping panel design in which any given sample panel is interviewed in-person a total of five times over 30 months to yield annual use and expenditure data for 2 calendar years. These rounds of interviewing are conducted at about 5- to 6-month intervals. They are administered through a computer-assisted personal interview mode of data collection and take place with a family respondent who reports for him/herself and for other family members. Data from two panels are combined to produce estimates for each calendar year.

For more information on MEPS data and the underlying survey methodology, please visit https://www.meps.ahrq.gov/mepsweb/.

3) *Why were some PDS clinical trial datasets linked with MEPS but others were not?*

PDS clinical trial datasets were generally linked with MEPS when they contained the required linking variables: age, race, sex, and a health-related quality of life measurement. Why these four variables? For most patient-level records on the PDS platform, demographic measures are limited to age, race, and sex to reduce the possibility of re-identification. A data integration effort limited to these three demographic measures would produce a multitude of many-to-many exact linkages. To ameliorate this problem, the linkage approach used an additional measure that further distinguishes patients by their health-related quality of life assessments. This measure is the EQ-5D™ index score, derived from the EuroQoL five dimensions questionnaire, one of the most commonly used measures of health-related quality of life. Additional demographic variables were integrated into the linkage process when available.

Although not a strict requirement for linkage, clinical trials that studied more prevalent cancers were also prioritized for linkage with MEPS, as these cancer types have greater representation among MEPS respondents.

4) *What types of analysis can be conducted using a PDS-MEPS linked dataset?*

The PDS-MEPS linked datasets enable researchers to study relationships between the appended MEPS variables (e.g., socio-economic, health, and health care use characteristics) and clinical trial outcomes of interest. Researchers can also conduct probabilistic assessments to understand whether the clinical trial is representative of socio-demographic subgroups in the U.S. population.

5) *What resources are available for new users of PDS-MEPS linked datasets?*

Example SAS Programs: For examples of data processing steps and basic analyses, users can reference the code package:
https://data.projectdatasphere.org/projectdatasphere/html/resources/PDF/MEPS_DATA_EXAMPLE

- Description of Linkage Methods: For details on the linkage methods used to construct a specific PDS-MEPS linked dataset, users can reference the documentation that accompanies the linked dataset of interest. A general discussion of the linkage methods can be found in Cohen & Unangst (2018), accessible at the following link. https://www.frontiersin.org/articles/10.3389/fonc.2018.00365/full

- Codebooks: A codebook is available for each linked dataset. Because some variables in each linked dataset originate from MEPS, we have also provided a crosswalk that explains variable recodes and collapsing schemes that were performed on the source MEPS data.

- Example Analyses: For examples of analysis conducted with a PDS-MEPS linked dataset, see the results section of Cohen & Unangst (2018) which can be accessed at https://www.frontiersin.org/articles/10.3389/fonc.2018.00365/full.

6) *Who created the linked datasets, and who funded the work?*

The linked datasets were created by RTI International in collaboration with Project Data Sphere®. RTI International is an independent nonprofit research institute headquartered in Research Triangle Park, NC. Please visit https://www.rti.org/ for more information.

Support for this work was provided by the Robert Wood Johnson Foundation, Grants # 74096 and 77003.