Enhancing Patient-Level Clinical Trial Data with Medical Expenditure Panel Survey Data: Quick Start Guide to the Enhanced Datasets

Jennifer Unangst[1], Steven B. Cohen[1], Feng Yu[1]
[1]RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Abstract
The Project Data Sphere® (PDS) online platform provides the cancer 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, but to address the confidentiality provisions inherent to the trials, data providers are required to mask or remove certain demographic data. The de-identification process limits researchers' 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. To overcome these analytic constraints, our team created a series of enhanced datasets, whereby content from the nationally representative Medical Expenditure Survey (MEPS) has been appended to patient-level data from select clinical trials. Comparator arm patients from the clinical trials were deterministically matched with similar cancer survivors from MEPS based on age, sex, race, and quality of life. In this paper, we describe the enhanced datasets, the types of analyses they support, and the free resources available for data users.

Key Words: Project Data Sphere; Survey data; Data integration, Clinical trial data

1. Introduction

Project Data Sphere® (PDS) is an open-access cancer research platform that launched in 2014 as an independent initiative of the CEO Roundtable on Cancer's Life Sciences Consortium. It hosts de-identified patient-level clinical trial data from more than 200 Phase III cancer clinical trials representing over 240,000 cancer patients. A primary goal of PDS is to unleash the full potential of existing clinical trial data and advance new research efforts that will improve the lives of cancer patients and their families around the world (Green et al., 2015). While these data are rich in measures that characterize the clinical trials under study, data providers are required to mask or remove key social and demographic data to preserve patient confidentiality. Consequently, researchers have limited ability to study the influence of health-related and socioeconomic factors on treatment effects and patient outcomes.

With support provided by the Robert Wood Johnson Foundation, PDS and RTI International collaborated to address these analytic constraints. For a selection of clinical trials on the 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 set of enhanced linked datasets, each 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. 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 of like cancer survivors in the U.S. population.

These enhanced datasets can now be freely downloaded from the PDS website. This paper provides a high-level introduction to the enhanced datasets. Section 2 summarizes the data sources and the methods used to integrate them. Section 3 describes the contents of the enhanced datasets and types of analyses they support. Section 4 highlights several resources available to aid new data users. Section 5 discusses limitations to consider when conducting analyses with the enhanced data sets.

2. Creation of the PDS-MEPS Enhanced Datasets

2.1 Medical Expenditure Panel Survey

The source of nationally representative survey data that was used to link with the clinical trial datasets was 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). For our purposes, we'll focus on the Household Component (MEPS-HC).

The MEPS-HC is an annual survey of roughly 14,000 households in the U.S. It 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.

The MEPS-HC is designed to provide national and state level estimates for the most populous states, covering a variety of health-related topics, such as health behaviors and perceptions, access to, use, and quality of health care, as well as health expenditures. An attractive feature of MEPS that makes it amenable to being integrated with the PDS clinical trial datasets is that it includes self-reported medical conditions data. Each respondent can enumerate medical conditions they've had, which are then coded into well-defined categories. These medical condition codes can be used to identify cancer survivors that have the same general type of cancer (e.g., breast cancer, prostate cancer) as a specific clinical trial dataset. Another-attractive feature is that each year of MEPS data is designed to represent the U.S. general population on its own. While a single year of MEPS may only contain, say 30 gastric cancer survivors, the MEPS data files can be pooled across years to boost representation of the cancer survivors of interest. This provides a richer set of survivors to potentially link with the patients represented in PDS clinical trials.

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

2.2 Clinical Trial Datasets from the PDS Platform

The data available on the PDS platform represent de-identified, patient-level, randomized clinical trial data. These data characterize the clinical trials under study, their treatment protocols, and patient outcomes. For each trial, data have typically been provided as a suite of SAS datasets, each featuring a topical area such as demographic, medical history, physical examination, medication, hospitalization, exposure, tumor assessment, cancer treatment, and death information, etc. These datasets can be studied individually or be combined for more comprehensive patient-level analyses. Details of the data for each cancer clinical trial can be found in the documentation provided for the trial on the PDS website.

As noted previously, for confidentiality purposes, some key social-demographic information of the patients enrolled in each study has been removed prior to release for public use. Our work focused on the clinical trial datasets for which the required linkage variables were available (see Section 2.3) and which represented types of cancer that were also observed in MEPS. The set of PDS clinical trial datasets that met these criteria only contained comparator arm patients. Consequently, only comparator arm patients were linked with MEPS.

2.3 Integration Approach

It is highly unlikely that the same individuals are represented in both the PDS clinical trial data and MEPS survey data. Instead, the goal of our work was to link similar PDS clinical trial patients and MEPS cancer survivors having the same type of cancer, so the MEPS cancer survivors can effectively donate their survey variables for analysis.

The linkage approach utilizes variables that are available in both data sources. For most patient-level records on the PDS platform, the demographic measures available for linkage are limited to age, race, and sex to reduce the possibility of patient re-identification. Using only these three variables would produce a multitude of linkages. Consequently, our linkage process incorporates an additional quality of life measure, called the EQ-5D, which distinguishes patients by their health-related quality of life assessments. The EQ-5D consists of the following five health-related components: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. These five components are collectively used to generate an overall quality of life score (Cohen and Unangst, 2018). When additional variables were available and common to both data sources, they were incorporated into the linkage process as well. For example, BMI category was used as an additional linkage discriminator for several of the trials we examined.

The linkage process itself consisted of multiple steps, where the linkage criteria were gradually relaxed across steps. For example, the first pass at forming linkages required exact matches on single year age, race, sex, and EQ-5D quality of life score. The second pass only required a match on categorized age, race, sex, and EQ-5D score. The third pass further relaxed the criteria by requiring an exact match on categorized age, race, sex, and categorized EQ-5D score. Thus, the linkages formed at earlier steps represent stronger quality matches. Many-to-many linkages were permitted in this process, so that if there were 2 PDS patients with the same profile, and there were 3 MEPS cancer survivors with that same profile, the process would yield 6 linkages.

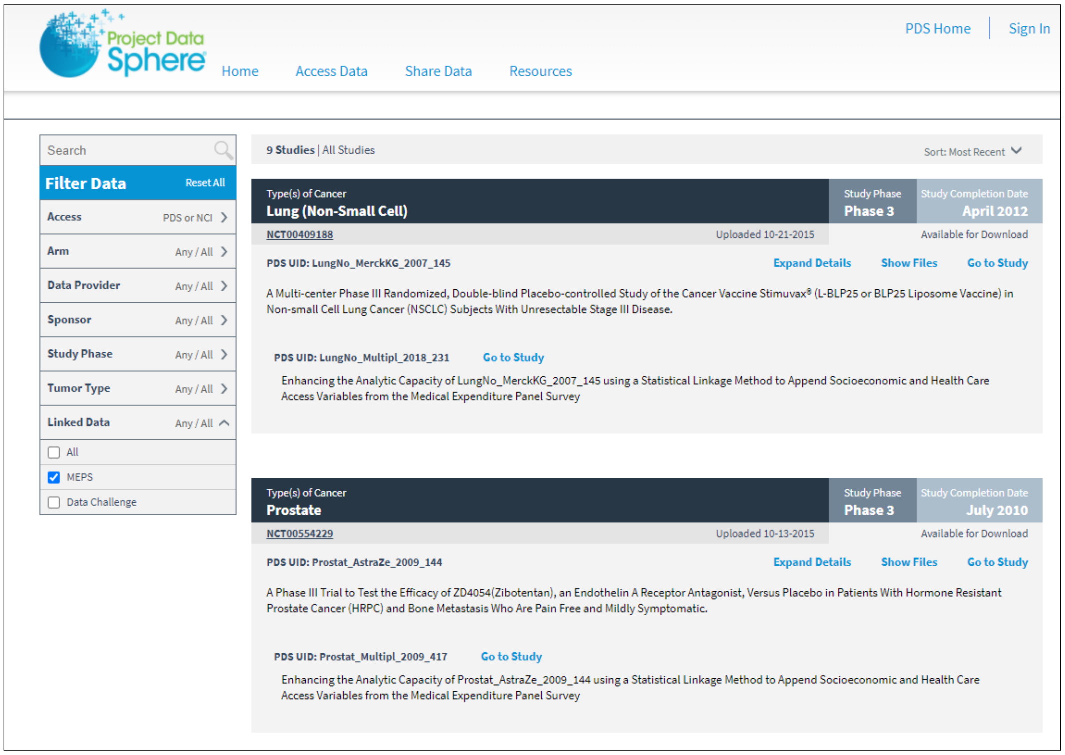
Data users are highly encouraged to review details of the linkage process prior to conducting analysis. For details on the linkage methods used to construct a specific PDS-MEPS linked dataset, users can reference the documentation that accompanies the linked dataset of interest. A more detailed discussion of the linkage methods can also be found in Cohen & Unangst (2018).

3. About the Enhanced Datasets

3.1 Accessing the Data

To access the enhanced analytic datasets, users can visit the PDS website and sign up for a free account if they don't already have one.
https://data.projectdatasphere.org/projectdatasphere/html/access

As shown in **Figure 1**, after navigating to the Access Data page, users can filter specifically to the enhanced datasets via the filtering menu on the left side of the screen. Searching for MEPS linked data will return nine enhanced datasets.



Access	PDS or NCI
Arm	Any / All
Data Provider	Any / All
Sponsor	Any / All
Study Phase	Any / All
Tumor Type	Any / All
Linked Data	Any / All

□ All
☑ MEPS
□ Data Challenge

9 Studies | All Studies

Type(s) of Cancer: **Lung (Non-Small Cell)** — Study Phase: **Phase 3** — Study Completion Date: **April 2012**

NCT00409188 — Uploaded 10-21-2015 — Available for Download

PDS UID: LungNo_MerckKG_2007_145

A Multi-center Phase III Randomized, Double-blind Placebo-controlled Study of the Cancer Vaccine Stimuvax® (L-BLP25 or BLP25 Liposome Vaccine) in Non-small Cell Lung Cancer (NSCLC) Subjects With Unresectable Stage III Disease.

PDS UID: LungNo_Multipl_2018_231

Enhancing the Analytic Capacity of LungNo_MerckKG_2007_145 using a Statistical Linkage Method to Append Socioeconomic and Health Care Access Variables from the Medical Expenditure Panel Survey

Type(s) of Cancer: **Prostate** — Study Phase: **Phase 3** — Study Completion Date: **July 2010**

NCT00554229 — Uploaded 10-13-2015 — Available for Download

PDS UID: Prostat_AstraZe_2009_144

A Phase III Trial to Test the Efficacy of ZD4054(Zibotentan), an Endothelin A Receptor Antagonist, Versus Placebo in Patients With Hormone Resistant Prostate Cancer (HRPC) and Bone Metastasis Who Are Pain Free and Mildly Symptomatic.

PDS UID: Prostat_Multipl_2009_417

Enhancing the Analytic Capacity of Prostat_AstraZe_2009_144 using a Statistical Linkage Method to Append Socioeconomic and Health Care Access Variables from the Medical Expenditure Panel Survey

Figure 1. Filtering Criteria to Access the PDS-MEPS Enhanced Datasets

The results returned when filtering to the MEPS linked datasets include the source data files from the clinical trial (i.e., the raw clinical trial datasets used to link with MEPS). Those raw files are presented alongside the corresponding enhanced dataset. Most researchers will likely want to download both data types. For efficiency purposes, the enhanced datasets do not include any substantive variables from the clinical trial (except the linkage variables). Researchers who wish to incorporate clinical trial variables into their analysis can append those variables by linking back to the source clinical trial datasets using the PDSID provided in the enhanced dataset.

3.2 Dataset Contents

Each enhanced analytic dataset contains two types of records. The first type of record (identified using variable LINKSTATUS=1 in the dataset) represents the many-to-many linkages, or pairs of patient IDs from the clinical trial of interest and cancer survivor IDs from MEPS. The second type of record (identified using variable LINKSTATUS=2 in the dataset) represents the MEPS cancer survivors that were eligible to link with PDS patients but did not form a linkage. This second group of records has been included in the enhanced dataset, because they can be helpful for assessments of representational disparities that examine the profiles of patients enrolled in the trials relative to the population of like cancer survivors in the U.S. population.

The variables included in each enhanced dataset include the following types. Capitalized words indicate the variable name in the enhanced dataset. As previously noted, the enhanced datasets do not include any substantive variables from the clinical trial of interest.

- Type of analytic record, as described above (LINKSTATUS)
- Set of criteria used to form the PDS-MEPS linkage (LINKMETHOD)
- Linkage variables (e.g., age, race, sex, items used to calculate EQ-5D score)
- Patient and cancer survivor IDs (PDSID and MEPSID, respectively)
- Selection of MEPS survey variables

The contents can be explored more fully using the supporting documentation provided with each enhanced dataset. These files include documentation of data processing and linkage methods, a codebook of the data contents, and a crosswalk between the variables provided in the enhanced dataset and their source.

3.3 Analyses Supported by the Enhanced Datasets

Presented next are two example analyses that can be conducted using the enhanced datasets. The examples presented were produced using the linked data for NCT00409188: A Multi-center Phase III Randomized, Double-blind Placebo-controlled Study of the Cancer Vaccine Stimuvax (L-BLP25 or BLP25 Liposome Vaccine) in Non-small Cell Lung Cancer (NSCLC) Subjects With Unresectable Stage III Disease.

3.3.1 Exploring Factors Potentially Associated with Survival

The first type of analysis is one that explores relationships between outcomes from the clinical trial and covariates from both the clinical trial and from MEPS. **Table 1** presents example output from a logistic regression model of overall survival status from the clinical trial. The independent variables in the model represent a mix of measures from the clinical trial, such as response to chemo-radiotherapy or cancer stage, and from MEPS such as income, insurance coverage, smoker status, beliefs about health insurance, and whether or not the person ever had lab tests. The results indicate that, in addition to the PDS measure reflecting the stage of the lung cancer tumor (N stage), a cancer patient's likelihood of survival was associated (P-value < 0.05) with their insurance coverage status and the intensity of services received in their ambulatory health care visits, as represented by having had lab tests (Table 1). Smoking status and patient health preferences were also found to be mildly associated with survival (P-value < 0.10). Additional examples of this type may be found in Cohen & Unangst (2018).

Table 1: Exploratory Assessment of Factors Suggesting Association with Lung Cancer Survivorship in the Comparator Arm

Contrast	DF	Wald F	P-value
Overall Model	11	7.11	<0.0001
Model minus intercept	10	3.49	0.0002
Clinical Trial Measures			
Response to chemo-radiotherapy	1	2.82	0.094
Type of chemo-radiotherapy	1	2.49	0.1155
N Stage	2	4.24	0.0151
MEPS Measures			
Income	1	2.33	0.1273
Medicaid coverage	1	7.43	0.0067
Private HMO coverage	1	3.34	0.0684
Smoker status	1	3.81	0.0515
Believes health insurance is not needed	1	3.64	0.0573
Had lab tests	1	6.05	0.0143

Data users must follow a few careful steps to conduct this kind of analysis. First, the analytic dataset should be filtered to records that represent linkages between PDS patients and MEPS cancer survivors (LINKSTATUS = 1). Second, because the enhanced datasets do not include measures from the clinical trial, the data user should append any clinical trial items of interest to the enhanced analytic datasets using the PDSID. Third, because the enhanced dataset includes many-to-many linkages, running an analysis where the unit of analysis is the clinical trial patient requires deduplicating by PDSID, or essentially choosing a MEPS cancer survivor to donate the values of MEPS variables for each patient. There are a number of ways to choose a MEPS donor for each patient, each of which has pros and cons. The LINKMETHOD variable may be helpful as data users consider their approach, because LINKMETHOD summarizes the criteria that were used to generate each linkage. Using LINKMETHOD, it's possible to distinguish stronger linkages from weaker ones. As a general word of caution for this type of analysis, results will be sensitive to which MEPS case is used as a donor. It is advised to run some sensitivity analyses to assess stability of the results. See our 2021 AAPOR proceedings paper Cohen, Unangst, & Yu (forthcoming) for a summary of sensitivity analyses conducted to date.

3.3.2 Example of Representational Disparities Assessment

Clinical trials are often conducted among younger, healthier, and less racially diverse patient populations than the population at large (Denson & Mahipal, 2014; Hamel et al., 2016; O'Keefe et al., 2015). Consequently, a second type of analysis that may be of interest is whether particular types of patients were over-/under-represented in the clinical trial's comparator arm compared to the U.S. population of like cancer survivors. This can be explored by examining associations between the characteristics of the MEPS cancer survivors and their linkage status with patients in the clinical trial of interest (i.e., are the MEPS cancer survivors that linked with a PDS patient different from those who didn't?).

Table 2 presents example output from a logistic regression where the dependent variable is linkage status (i.e., whether or not the MEPS cancer survivor linked with a PDS patient in the clinical trial of interest), and the independent variables are items from MEPS. Based on the results of the logistic model, the following measures were identified as significant predictors (P-value < .05) of having a greater likelihood of being represented in the trial: race/ethnicity, sex, marital status, MEPS survey year, EQ-5D, and smoker status. More specifically, the lung cancer patients enrolled in the trial were more likely to be men, white, married, and current smokers relative to their representation in the population. Individuals characterized by fewer health problems as noted by higher values of the EQ-5D were also more likely to be enrolled in the trial. The inclusion of the MEPS survey year variable was a methodological consideration, serving to control for the estimation strategy utilized for the EQ-5D measurement. For additional examples of representational assessments conducted with the enhanced datasets, please see Cohen, Unangst, & Yu (2020).

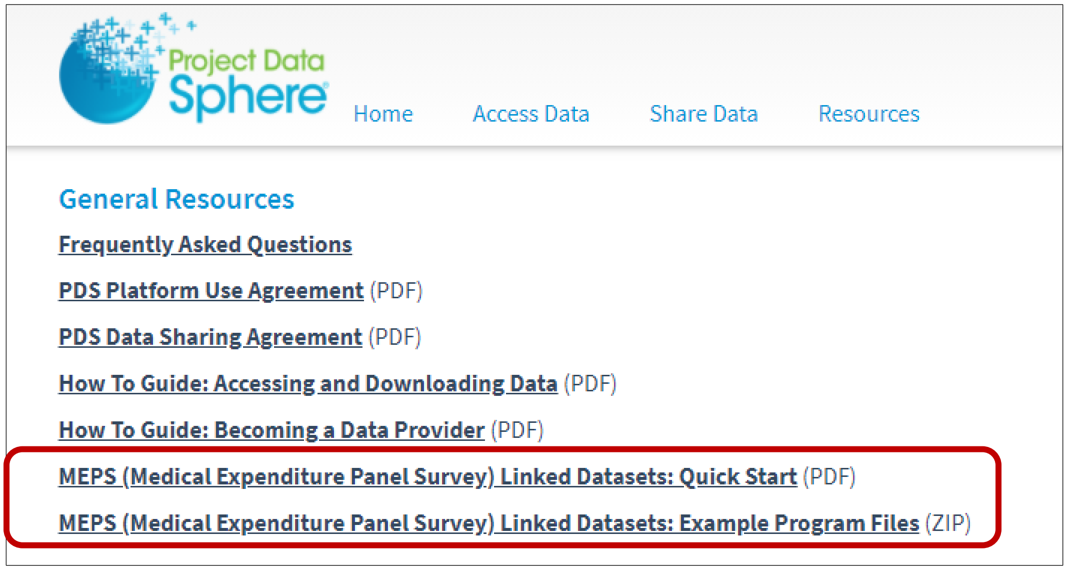
Table 2: Factors that Distinguished the Characteristics of Lung Cancer Patients Enrolled in the PDS Clinical Trial

Contrast	DF	Wald F	P-value
Overall Model	9	6.95	<0.0001
Model minus intercept	8	7.1	<0.0001
Marital Status	1	6.16	0.0134
Sex	1	11.04	0.0010
MEPS Year	1	6.47	0.0113
EQ5D Decile Category	1	14.81	0.0001
Race/Ethnicity	2	25.94	<0.0001
Difficulty in access to necessary medical care	1	3.17	0.0758
Smoker Status	1	4.46	0.0352

To conduct this kind of analysis, data users should first deduplicate the enhanced analytic dataset by the MEPSID, so each MEPS cancer survivor is represented only once. Each MEPS case has a linkage status assigned to it (LINKSTATUS), which summarizes whether it linked with a PDS patient for the clinical trial of interest or not. Using the linkage status indicator, it's then possible to run assessments that compare the profiles of the linked and unlinked MEPS cases to understand which types of individuals were more likely to be represented in the comparator arm of the trial. When running analyses that utilize MEPS cases as the units of analysis, the data user should specify the MEPS complex sample design to the statistical software being used. Because the MEPS data were collected using a complex sample design, it's not appropriate to analyze the data as if they were generated from a simple random sample (the default for most statistical software unless you tell it otherwise). To describe the design to the statistical software, the user will provide the analysis weight, variance strata, and variance cluster indicators to the software. For examples of how to specify the MEPS design when using SAS, SUDAAN, STATA, and SPSS, please see Machlin, Yu, and Zodet (2005) linked below.
https://meps.ahrq.gov/mepsweb/survey_comp/standard_errors.jsp

4. Resources Available for Data Users

Provided with each enhanced dataset is a suite of documentation files, including a codebook, methodological documentation describing the linkage procedures for the specific trial of interest, and a variable crosswalk that maps the contents of the enhanced dataset back to corresponding variables in the source data files (e.g., MEPS public use files or clinical trial datasets). Users can also visit the Resources page on the PDS website (https://data.projectdatasphere.org/projectdatasphere/html/resources) for more general resources, such as a Quick Start Guide that introduces the enhanced datasets and sample SAS programs that demonstrate how to process the enhanced datasets and run simple analyses (see **Figure 2**).



General Resources

Frequently Asked Questions

PDS Platform Use Agreement (PDF)

PDS Data Sharing Agreement (PDF)

How To Guide: Accessing and Downloading Data (PDF)

How To Guide: Becoming a Data Provider (PDF)

MEPS (Medical Expenditure Panel Survey) Linked Datasets: Quick Start (PDF)

MEPS (Medical Expenditure Panel Survey) Linked Datasets: Example Program Files (ZIP)

Figure 2. Locating Additional Resources for the Enhanced Datasets on PDS Website

5. Limitations

There are a few limitations worth noting about the PDS-MEPS enhanced datasets. First, enhanced datasets are only available for a subset of clinical trials available on the PDS platform. The trials that have been linked with MEPS are limited to those containing the required linkage variables and that represented types of cancer that are also observed among MEPS respondents. Second, when interpreting results from analyses run on the linked datasets, data users should be mindful of representational differences between the contributing clinical trial patients and MEPS respondents. For example, MEPS cases can represent respondents from 2000-2016, while the clinical trials may have been conducted in a different time frame. The MEPS cases also represent cancer survivors in the U.S., while the clinical trials usually represent patients from around the world. Lastly, when data users conduct analyses that use variables from both the clinical trial of interest and from MEPS, variance estimates produced by the statistical software may be somewhat underestimated, because it is non-trivial to produce variance estimates that account for the different underlying designs of MEPS and the clinical trials.

6. Concluding Remarks

The PDS-MEPS enhanced 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 of like cancer survivors in the U.S. population. Prior to beginning analysis with one of the linked datasets, users are highly encouraged to explore existing publications containing example analyses and the resources available from the PDS website.

Acknowledgements

This effort reflects a collaboration between RTI International, Project Data Sphere, LLC, and the Robert Wood Johnson Foundation. Special acknowledgements go to Dr. Katherine Hempstead, Robert Wood Johnson Foundation; Dr. Martin Murphy, Dr. William Louv, and Mr. Michael Litzinger, Project Data Sphere, LLC; and Dr. Georgiy Bobashev, RTI International, for their contributions. This research was made possible through a grant from the Robert Wood Johnson Foundation. Grant # 77003. The views expressed here do not necessarily reflect the views of the Foundation.

References

Cohen S. B., Unangst J., Yu F. (forthcoming). Reproducibility assessments of analytic findings derived through national survey data integration efforts. 2021 Proceedings of the American Association of Public Opinion Research.

Cohen, S. B., Unangst, J., & Yu, F. (2020). Enhancing the analytic utility of clinical trial data to inform health disparities research. *Contemporary Clinical Trials Communications*, *20*. https://doi.org/10.1016/j.conctc.2020.100677

Cohen S. B., Unangst J. (2018). Data integration innovations to enhance analytic utility of clinical trial content to inform health disparities research. *Frontiers in Oncology*, 8, 365.

Denson, A. C., & Mahipal, A. (2014). Participation of the elderly population in clinical trials: Barriers and solutions. *Cancer Control,* 21(3), 209–214. http://dx.doi.org/10.1177/107327481402100305

Green A. K., Reeder-Hayes K. E., Corty R. W., et al. (2015). The Project Data Sphere Initiative: Accelerating Cancer Research by Sharing Data. *The Oncologist.* 20(5): 464–e20. http://dx.doi.org/10.1634/theoncologist.2014-0431

Hamel, L. M., Penner, L. A., Albrecht, T. L., Heath, E., Gwede, C. K., & Eggly, S. (2016). Barriers to clinical trial enrollment in racial and ethnic minority patients with cancer. *Cancer Control,* 23(4), 327–337. http://dx.doi.org/10.1177/107327481602300404

Machlin, S., Yu, W., and Zodet, M. (2005). Computing Standard Errors for MEPS Estimates. Agency for Healthcare Research and Quality, Rockville, MD. https://meps.ahrq.gov/mepsweb/survey_comp/standard_errors.jsp

O'Keefe, E. B., Meltzer, J. P., & Bethea, T. N. (2015). Health disparities and cancer: Racial disparities in cancer mortality in the United States, 2000-2010. *Frontiers in Public Health*, 3, 51. http://dx.doi.org/10.3389/fpubh.2015.00051