Assessing Reproducibility of Analytic Findings Derived through National Survey Data Integration Efforts: A Case Study Linking Patient-level Clinical Trial Data with Medical Expenditure Panel Survey Data

Steven B. Cohen[1], Jennifer Unangst[2], Feng Yu[3]
[1, 2, 3]RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Abstract

The quality and content of national population-based surveys are enhanced through integrated designs that link additional medical, behavioral, environmental, socio-economic and financial content from multiple sectors. A recent effort by the Committee on National Statistics of the National Academy of Sciences is serving as a catalyst to advance future national data integration efforts, as indicated in their recent report on Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. These integrated data platforms include content drawn from administrative sources and research studies to enhance analytic capacity.

The integration of national survey data with content derived from multiple data sources has the capacity to provide greater insights than possible from any of the component sources. Based upon the level of granularity characterizing the data profiles of the component data sets, the data integration effort can result in many-to-many linkages. While it is preferential to restrict these data enhancement efforts to situations that permit one-to-one linkages, this is often not possible. In this paper, we focus on a data integration effort that is impacted by many-to-many linkages and provide a model for assessing the reproducibility of analytic findings attributable to alternative linkage applications. Examples are provided using data from the Medical Expenditure Panel Survey (MEPS) and cancer patient-level phase III clinical datasets. This data integration effort permits studies assessing the influence of health-related and socioeconomic factors, access to and use of health care services, health behaviors and preferences in concert with clinical trial treatment effects on cancer patient outcomes, heretofore not possible. In this study, we assess the reproducibility and stability of the analytic findings identifying factors influencing patient outcomes as the linkages are modified. Building similar evaluations into data integration efforts may serve to provide additional evidence in support of the integrity resultant findings.

Key Words: Project Data Sphere; Data Integration; Health Disparities; MEPS; Reproducibility

1. Background

The *Project Data Sphere*® (PDS) online platform is a centralized place where the cancer research community can broadly share, integrate, and analyze historical patient-level data from academic and industry phase III clinical trials. 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 PDS data are rich in measures that characterize the clinical trials under study, data providers are required to de-identify patient-level data for patients' confidentiality protection by removing key social and demographic content that could otherwise be used to study underserved populations and the complex social, behavioral, and biological factors that contribute to inequities. To address these analytic constraints, with support provided

by the Robert Wood Johnson Foundation, PDS and RTI International are collaborating to enhance the analytical utility of selected PDS datasets (downloadable from www.ProjectDataSphere.org). The effort has augmented the data profiles of cancer patients in selected PDS clinical trial datasets with social, economic, and health-related content from the nationally representative Medical Expenditure Panel Survey (MEPS). Patients from a representative set of PDS clinical trials were statistically linked with similar cancer survivors from MEPS to append measures of health care access and utilization, patient behaviors and attitudes toward care, and health conditions. This collection of content-enhanced PDS resources permit researchers to conduct probabilistic assessments of the representativeness of the cancer patients in these trials, and identify health disparities impacting on health outcomes. This initiative has been advanced to achieve the following objectives:

- *To broaden the analytic capacity of PDS clinical trial data in support of health disparities and health outcomes research for cancer patients*;
- *To significantly scale up the analytic utility and content that can be realized by these data integration efforts*;
- *To conduct a broad array of assessments that investigate the representativeness of cancer clinical trial patients relative to characteristics of cancer survivors in the U.S. general population*.

Linking the PDS-MEPS data resources also enable more targeted analyses that examine questions such as: How do disparities in cancer patients' access to health care and income impact patient outcomes in specific phase III clinical trials? What variations in patient outcomes are associated with specific demographic, socioeconomic, and health-related factors?

Project Data Sphere, LLC (PDS), an independent, not-for-profit initiative of the *CEO Roundtable on Cancer's Life Sciences Consortium* (LSC), operates the *Project Data Sphere* platform, a free digital library-laboratory where the research community can broadly share, integrate and analyze historical, patient-level data from academic and industry phase III cancer clinical trials. PDS hosts over 200 phase III oncology clinical trial datasets, representing more than 150,000 cancer patients. This initiative extends the utility of these data by joining PDS patient-level data with nationally representative health-related data from the Medical Expenditure Panel Survey (MEPS). MEPS, sponsored by the Agency for Healthcare Research and Quality (AHRQ), is the nation's primary source of nationally representative, comprehensive, person-level data on health care use, insurance coverage, and expenses. Over the past several years, the MEPS data have supported a highly visible set of descriptive and behavioral analyses of the U.S. health care system (Cohen and Cohen, 2013).

Using data integration methods, sociodemographic, access, health, and health care-related measures associated with a nationally representative set of cancer survivors from MEPS are linked to similar cancer patients in the PDS analytic datasets using variables available in both data sources -- demographic information (age, race/ethnicity, and sex) and the EQ-5D™ index score, derived from the EuroQoL five-dimensions questionnaire (Cohen and Unangst, 2018). When additional demographic measures are available in both datasets (e.g., body-mass index), they are also incorporated in the linkage process.
The MEPS typically surveys 2,000 participating sample adults aged 18 and older with a reported cancer diagnosis. Several years of MEPS data on cancer survivors may be pooled to enhance the sample sizes of cases available for specific cancer classifications;

this results in a much larger set of survivors of various cancer types available for linkage. The MEPS data files are accessible for downloading at the MEPS website: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

2. Assessing the Sensitivity of Patient-level Analyses to MEPS Donor Selection

Each PDS-MEPS linked dataset facilitates three types of analysis. First, data users can produce nationally representative estimates of cancer survivors in the U.S. who have the same general cancer type (e.g., breast cancer, prostate cancer) as the individuals in the clinical trial. Second, data users can assess representational disparities between the patients enrolled in the comparator arm of the clinical trial and the general population of U.S. cancer survivors with the same general type of cancer; see Cohen, Unangst, & Yu (2020) for an example of this type of representational assessment. These first two types of analysis are primarily driven by the sociodemographic, healthcare, and health-related content from MEPS as well as information on the linkage process itself (e.g., whether or not certain MEPS cancer survivors were linked with any patients from the trial of interest can indicate inclusion or exclusion of particular patient profiles from the trial). For the third type of analysis, data users may wish to analyze patient-level outcomes from the clinical trial and incorporate the sociodemographic and health-related content appended from MEPS. This last type of analysis requires greater care and a clearer understanding of the linkage process itself as described in Cohen & Unangst (2018).

Because the linkage process was deterministic, results from the first two types of analysis can be easily reproduced. To conduct the third type of analysis – that is, patient-level analysis using both variables from the clinical trial and from MEPS -- it is first necessary to account for the many-to-many linkages that resulted from the match process, as each PDS patient may have had multiple MEPS cancer survivors linked to it. One way to account for this is to select a single MEPS linkage for each PDS patient, so the selected MEPS cancer survivor can donate its survey variable values to that patient. Because most PDS patients were matched with more than one MEPS cancer survivor, the selection of a MEPS donor introduces some uncertainty into any analysis findings that utilize this approach. A further consideration is that, for any given PDS patient, linkages with MEPS may have been formed under different steps in the linkage process, where earlier steps involved stricter match criteria than later steps. Thus, there is additional variation introduced due to the varying quality of linkages across all MEPS cases associated with a PDS patient. Researchers using the PDS-MEPS linked datasets to conduct patient-level analyses that utilize MEPS content must consider these sources of variation carefully and would benefit from assessing sensitivity of results to the MEPS donor set. In this case study, we summarize our findings from several such sensitivity analyses. We address the following two research questions.

- Among PDS patients that were linked with MEPS, how sensitive are their MEPS-based representational distributions to the selection of MEPS donors?
- How sensitive are models of survival status to the selection of MEPS donors? How much do the factors identified as possible predictors of survival change as different MEPS donors are used in the analysis?

3. Methods

To assess the sensitivity of patient-level analyses to the selection of MEPS donors, we ran our patient-level analyses 200 times, where each iteration utilized a different set of donors. For iterations 1-100, we randomly selected one MEPS donor for each PDS patient from the set of its MEPS linkages formed under the strictest criteria that resulted in a match. For example, if a PDS patient formed four MEPS linkages under the strictest set of linkage criteria, we randomly chose one of those four MEPS cases to donate its values for the socio-demographic and health-related variables. In iterations 101-200, we randomly selected one MEPS donor for each PDS patient from the set of all MEPS cases linked to that patient. Thus, iterations 1-100 of the simulation prioritize the higher quality linkages when selecting a MEPS donor, while iterations 101-200 ignore differential linkage quality across MEPS matches.

We conducted the sensitivity assessment for two types of patient-level analyses that are possible with the linked datasets.

1. The first type of analysis looks at the socio-demographic and health-related characteristics of PDS patients that obtained a linkage with MEPS. While the patient characteristics derived from the clinical trial are stable and do not change based on MEPS donor, the characteristics appended from MEPS can change depending on donor selection. This has implications for whom is considered to be represented in a patient-level analysis.
2. The second type of analysis uses a logistic regression model to identify factors associated with survival outcome. Again, predictors derived from the clinical trial are fixed and do not change depending on MEPS donor; however, any socio-demographic or health-related predictors appended from MEPS will vary based on the donor. Thus, MEPS donors can influence which candidate predictors are deemed significant, as well as the direction of their relationship with survival outcome (e.g., they could either improve or reduce probability of survival).

To demonstrate our process, we present results for the clinical trials shown in Table 1. These trials were chosen as examples, because they produced the largest number of PDS-MEPS linkages among the more prevalent types of cancer. Given their larger number of linkages, results are expected to be more variable, and thus, more sensitive to selection of MEPS donors. All of these trials were linked with the 2000–2016 pooled MEPS data following the methods described in Cohen & Unangst (2018). PDS patients and MEPS cancer survivors were linked if they had an exact match by age, sex, race/ethnicity, EQ-5D quality of life score, and BMI category. A multi-step linkage process was used such that earlier linkage steps required stricter match criteria than later steps (e.g., earlier steps in the linkage process required an exact match by single-year age versus categorized age, or exact matches by a single-value of the EQ-5D score versus by categorized EQ-5D scores). Therefore, earlier linkage steps are expected to produce higher quality linkages than later steps. Linkage summaries for the trials can be found in Table 2.

Table 1. Clinical Trials Used as Examples of the Assessment Approach		
ClinicalTrials.gov Identifier	**Title**	**Unique PDS ID**
NCT00626548	A Phase III, Randomised, Placebo-controlled, Double-blind Study to Assess the Efficacy and Safety of Once-daily Orally Administered ZD4054 (Zibotentan) 10 mg in Non-metastatic Hormone-resistant Prostate Cancer Patients	Prostat_AstraZe_2008_103
NCT00081796	A Randomized, Open-Label, Phase III Study of RPR109881 IV Every 3 Weeks Versus Capecitabine (Xeloda) Tablets Twice Daily for 2 Weeks in 3-Week Cycles in Patients With Metastatic Breast Cancer Progressing After Taxanes and Anthracycline Therapy	Breast_SanofiU_2004_135

As shown in Table 2, linking the 677 comparator arm patients from Prostat_AstraZe_2008_103 with the 2,207 prostate cancer survivors surveyed in the 2000–2016 MEPS resulted in 11,694 linkages. Of the 677 comparator arm patients, 79% achieved at least one linkage with a MEPS cancer survivor. The median number of MEPS cancer survivors linked per PDS patient (among PDS patients with at least one linkage) was 12, with a range from 1-76. Among the 2,207 MEPS prostate cancer survivors that were eligible for linkage, 63% achieved at least one linkage with PDS. Of those, the median number of PDS patients linked per MEPS case was 6 with a range from 1-31. Linkage results for the Breast_SanofiU_2004_135 trial can be interpreted similarly.

Table 2. Linkage Summary Statistics for Selected PDS Clinical Trials		
	Prostate Cancer Trial 103	**Breast Cancer Trial 135**
Total Number of Linkages	11,694	2,392
Total PDS Patients	677	217
PDS Patients that Achieved at Least 1 Linkage	535 (79%)	122 (56%)
Number of MEPS Cases Linked per PDS Patient		
Median	12	10
Range (Min - Max)	1-76	2-156
Total MEPS Cases	2,207	2,987
MEPS Cases that Achieved at Least 1 Linkage	1,392 (63%)	1,448 (48%)
Number of PDS Patients Linked per MEPS Case		
Median	6	1
Range (Min - Max)	1-31	1-4

3.1 Characteristics of PDS Patients that Linked with MEPS: How Sensitive are their Representational Distributions to Selection of MEPS Donors?

Our first set of assessments examines the degree to which the MEPS-based representational distributions of PDS patients shift under different sets of MEPS donors. In other words, how much do the means of MEPS-based characteristics vary across MEPS donor sets?

For each trial serving as an example, Figure 1 presents a selection of the MEPS-based variables that we examined for PDS patients with at least one MEPS linkage. It includes several sociodemographic variables such as marital status, educational attainment, income, and U.S. Census region, as well as several health-related items including smoking status, poor health status, and having only public or any type of private insurance. For each variable, we present a smoothed distribution of means across all 200 simulated donor sets. We distinguish the first 100 donor sets from the second 100 donor sets to reflect the difference in donor selection method. Recall, the first 100 simulated donor sets were drawn by randomly selecting one MEPS donor for each linked PDS patient from the set of its MEPS linkages obtained under the strictest matching criteria (labeled as "strict" in the figure); the latter 100 simulated donor sets were drawn by randomly selecting one MEPS donor for each PDS patients from the set of its linkages obtained from any of the matching criteria (labeled as "any" in the figure).

Figures 1a and 1b present the distributional assessments for the prostate cancer trial and breast cancer trial, respectively. A few notable findings emerge from these figures.

- First, we can see that there is clearly variation across the mean estimates presented for each variable, which suggests that MEPS-based characteristics are indeed influenced by donor selection. This appears to be true for both trials we examined, although the degree of variability across donor sets differs by variable and by donor selection method. In general, the donor value distribution from either selection method seemed to have greater spread for the breast cancer trial than the prostate cancer trial.
- Second, while we had hypothesized that selecting a donor completely at random from all of a PDS patient's linkages would produce greater variability in MEPS-based characteristics than selecting a donor from the set of linkages achieved under the strictest criteria, this only held true for the breast cancer trial. As shown in Figure 1b, for the breast cancer trial, the "strict" donor selection method generally produced a distribution with thinner tails than the "any" donor selection method. However, in the prostate cancer trial, this trend was less consistent across variables. Some variables demonstrated a similar degree of variability across both donor selection methods (e.g., current smoker, private insurance), while others demonstrated less variability for the "any" selection method (e.g. income), which contradicted our hypothesis.
- Third, for most variables, the center of the distributions for the "any" and "strict" donor selection methods were generally fairly close to one another (within 10 percentage points). This suggests that, in expectation, selecting a donor under either approach should produce a roughly similar set of MEPS-based characteristics. A few exceptions occurred for the prostate cancer trial, where the distributions clearly diverged between the "strict" and "any" methods (e.g., private insurance, Census region).

Figure 1a. Distribution of Means for MEPS-based Characteristics of Linked PDS Patients

Summarized across 100 simulated donor sets per donor selection method

Prostate Cancer Trial 103

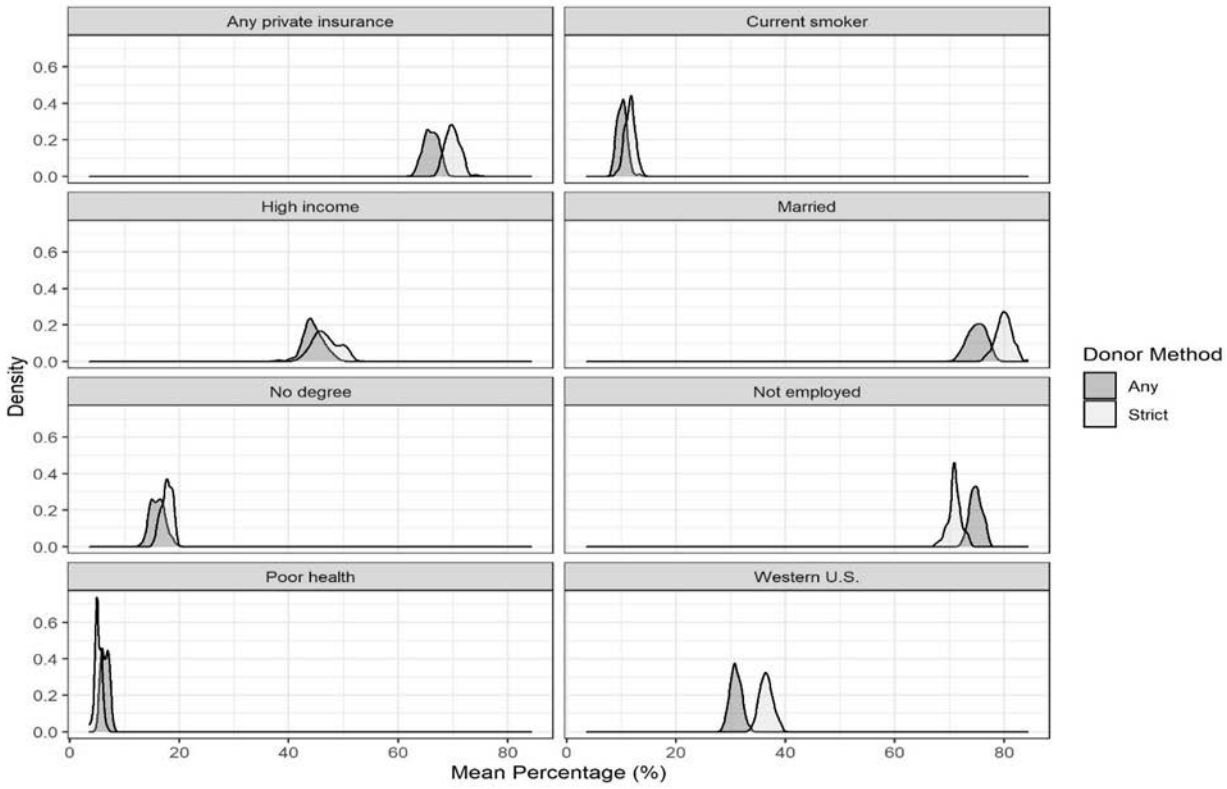
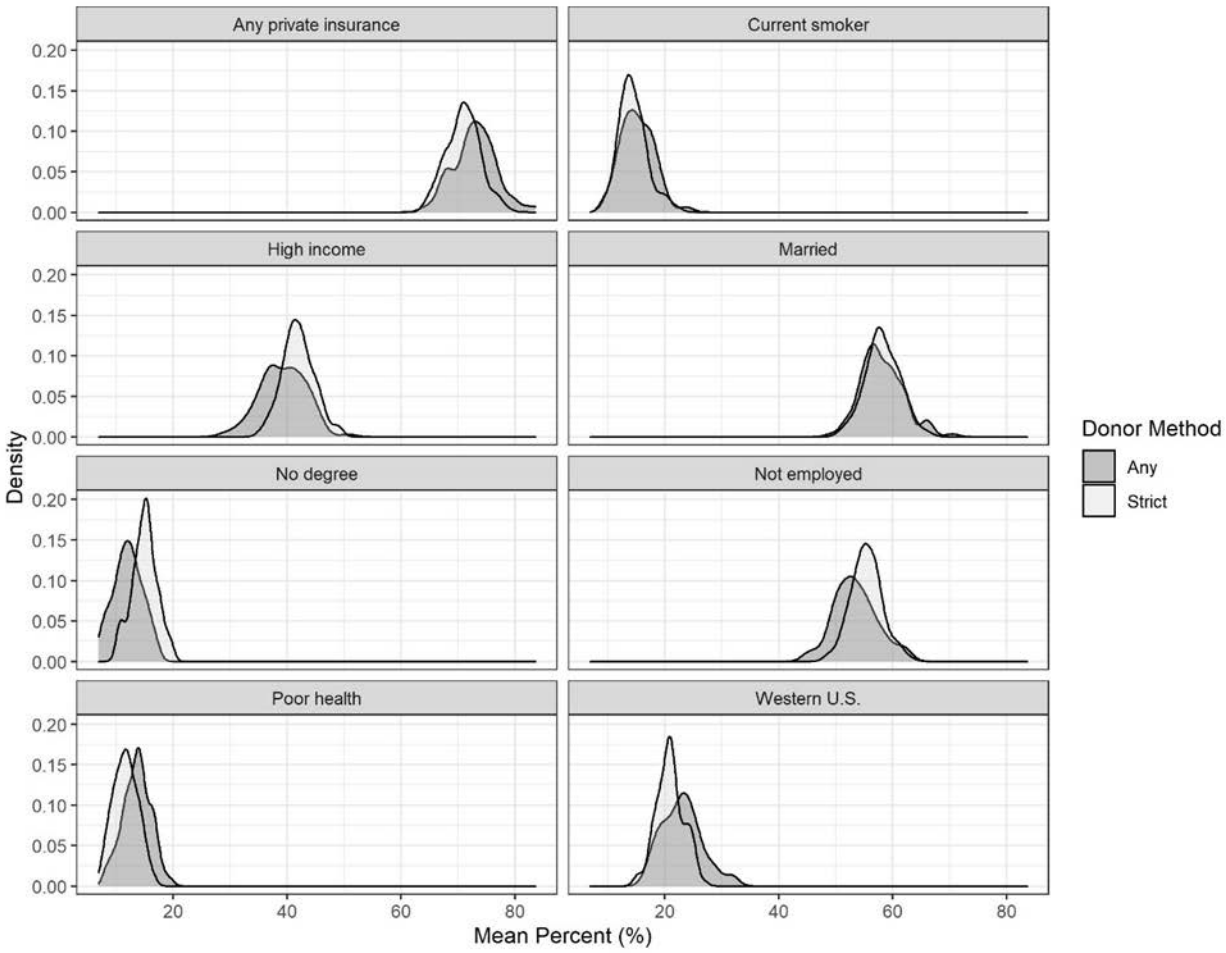


Figure 1b. Distribution of Means for MEPS-based Characteristics of Linked PDS Patients

Summarized across 100 simulated donor sets per donor selection method

Breast Cancer Trial 135



4. Factors Associated with Survival: How Sensitive are the Models to Selection of MEPS Donors?

Our second set of assessments examined the extent to which directional relationships and significance of predictors changed in logistic regression models of survival outcome. Tables 3 and 4 summarize the directional relationships and significance indicators, respectively, for candidate predictors across the 200 simulated donor sets; again, there are 100 donor sets for each of the two donor selection methods. For each trial, the same main effects model was fit across all donor sets to maximize comparability. Bivariate associations were manually reviewed for three randomly chosen donor sets to identify candidate predictors for each trial. The final set of candidate predictors for each trial generally included all linkage variables plus any variables that demonstrated a mild bivariate association with survival (p-value < 0.15) in at least one of the manually

reviewed donor sets. The source of each predictor is indicated in Tables 3a and 3b as either a linkage variable (common to both the PDS trial and MEPS), a MEPS variable from the survey, or a PDS variable from the clinical trial.

Tables 3a and 3b present the assessments of directional relationships between candidate predictors and survival outcome for the prostate cancer trial and breast cancer trial, respectively. To assess stability of directional relationships, these tables summarize the percentage of positive beta coefficients across the simulated donor sets for each variable in the logistic model. The numerator in the percentage is the number of times the beta coefficient was greater than zero, and the denominator is 100 (i.e., the total number of simulated donor sets used per selection method). Percentages closer to either 0% or 100% indicate that the variable's relationship with survival was highly consistent across donor sets, while percentages closer to 50% indicate fewer stable trends that are more sensitive to the selection of MEPS donors.

We observe several trends from these tables.

- Directional relationships between PDS-based predictors from the clinical trial and survival outcome were highly stable regardless of variability in either the linkage variables or MEPS-based predictors. In both trials we examined, the beta coefficients were consistently negative across all 200 donor sets.

- Directional relationships for linkage variables were moderately stable across donor selections. Interestingly, in the prostate cancer trial, we did observe one variable where the relationship appeared to reverse directions between the "strict" and "any" donor selection methods: age. Being age 65+ appeared to have a negative relationship with survival (i.e., reduced odds of survival) for the "strict" method with only 20% of the 100 donor sets having a positive beta coefficient. In contrast, being age 65+ demonstrated a slightly more positive relationship with survival (62% of beta coefficients were positive) for the "any" method. We note, however, that age was never found to be a significant predictor across all 200 donor sets we examined in the prostate cancer trial.

Table 3a. Percentage of Positive Beta Coefficients for Predictors of Survival Status

Summarized across 100 simulated donor sets per donor selection method

Prostate Cancer Trial 103

		Donor Selection Method	
		Strict	**Any**
	Intercept	100	100
Linkage	Age 65+	20	62
	Race, non-white	12	0
	EQ-5D category	74	96
	Overweight (BMI)	0	0
	MEPS donor from 2004-2016	100	91

MEPS	High income	58	50
	Any private insurance coverage	58	58
	No educational degree	59	75
PDS	Used opiates during trial	0	0
	Had new lesions during trial	0	0
	Had serious adverse events	0	0

Sources: Medical Expenditure Panel Survey Household Component, 2000–2016, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; PDS data file *Prostat_AstraZe_2008_103*, Project Data Sphere.

Table 3b. Percentage of Positive Beta Coefficients for Predictors of Survival Status

Summarized across 100 simulated donor sets per donor selection method

Breast Cancer Trial 135

		Donor Selection Method	
		Strict	**Any**
	Intercept	97	89
Linkage	Age 65+	8	25
	Race, non-white	0	0
	Overweight (BMI)	98	99
	EQ-5D category	100	98
	MEPS donor from 2004-2016	14	34
MEPS	Poor health	27	36
	Limited physical functioning	42	52
	Health insurance not worth the cost	29	36
	Pain limits work moderately/severely	56	53
	Unemployed	86	70
	Any private insurance coverage	34	25
PDS	Any new lesions during trial	0	0
	Overall tumor response: progressive disease	0	0

Sources: Medical Expenditure Panel Survey Household Component, 2000–2016, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; PDS data file *Breast_SanofiU_2004_135*, Project Data Sphere.

- Directional relationships for MEPS-based predictors were less stable than either the PDS-based predictors or linkage variables across donor selections, as might be expected. For the prostate cancer trial, the predictors were relatively unstable with a percentage of positive beta coefficients ranging from 58-59% for the "strict" donor selection method and 50-75% for the "any" donor selection method. For the breast cancer trial, some MEPS-based predictors demonstrated greater stability. For example, having poor health was clearly more likely to have a negative relationship with survival (27% positive beta coefficients) for the "strict" selection method.

Next, we examined the significance of Wald F effect tests for candidate predictors used in the prostate cancer and breast cancer models, respectively. To assess stability of statistical significance for each variable, the distribution of test results for each variable was assessed across four mutually exclusive significance levels: not significant, mildly significant (p-value < 0.10), moderately significant (p-value < 0.01), and highly significant (p-value < 0.001). For both trials, the results of significance testing were fairly stable across donor sets and selection methods. Generally, if a variable was not found to be significant at alpha level 0.10, this was the case across most or nearly all donor sets for a given method. Further, significance trends held across both donor selection methods. When variables did demonstrate less stability, results were still generally in the same realm of significance. For example, for the prostate cancer trial, the PDS-based variable "Had serious adverse events" was mildly significant 15% of the time, moderately significant 73% of the time, and highly significant 12% of the time. In this case, even though the precise significance level varied across donor sets, most models would still lead a researcher to believe that this variable has at least a moderate association with survival outcome.

5. Concluding Remarks

In summary, the PDS-MEPS linked datasets provide several analysis options for data users. In this report, we have focused on patient-level analyses that utilize sociodemographic and health-related survey content from MEPS, as appended through linkage. Because most PDS patients were matched with more than one MEPS cancer survivor, the selection of a MEPS donor for each patient introduces some uncertainty into results from patient-level analyses. To investigate the degree of sensitivity to MEPS donors, we conducted a simulation study for two types of patient-level analyses: (1) producing overall means of MEPS-based variables, (2) fitting models of survival outcome that incorporate MEPS-based variables as predictors.

Using the linked datasets for two prevalent cancers (i.e., prostate and breast cancers), we found that overall donor value distributions were influenced by selection of donors under both strict and relaxed selection criteria. The degree of variation in the means of MEPS-based variables depended on the variable and clinical trial. Using the breast cancer trial, for example, the centrality of donor value distributions tended to be more similar across strict and relaxed selection criteria than in the prostate cancer trial; however, the breast cancer distributions tended to demonstrate greater spread than the prostate cancer trial, on average. This could be related to differences in the number of potential donors per patient in each trial. While the median number of potential MEPS donors per PDS patient was similar across both trials (12 for the prostate cancer trial; 10 for the breast cancer trial), the maximum number of possible MEPS donors was much larger for the breast cancer trial compared to the prostate cancer trial.

In logistic regression models of survival outcome, we also observed that directional relationships of candidate predictors were somewhat sensitive to selection of MEPS donors. Directional relationships for PDS-based predictors (i.e., variables originating from the clinical trials) tended to be stable across donor selections, regardless of whether strict or relaxed donor selection criteria were used. Directional relationships for MEPS-based predictors were less consistent, however, across donor selections. For both trials, the results of significance testing were fairly stable across donor sets and selection methods. Generally, if a variable was not found to be significant at alpha level 0.10, for

example, this was the case across most or nearly all donor sets for a given donor selection method.

Overall, these findings suggest that using MEPS-based content in patient-level analyses should undergo reproducibility assessments. The enhanced MEPS-PDS data sets are well suited for exploratory assessments that primarily attempt to identify factors potentially associated with clinical trial outcomes. The stability regarding the direction of associations between predictors and survival outcomes are subject to level of consistency across donor selections. Due to the variability that results from donor selection, users of the linked datasets are cautioned to carefully examine findings from patient-level analyses and to triangulate results using other data sources, as possible.

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

Arfè A, Ventz S, Trippa L. Shared and Usable Data From Phase 1 Oncology Trials—An Unmet Need. *JAMA Oncol.* Published online June 04, 2020. doi:10.1001/jamaoncol.2020.0144

Cancer Health Disparities Research. National Cancer Institute website. https://www.cancer.gov/research/areas/disparities Updated December 19, 2018. Accessed June 11, 2020

Cohen SB, Cohen JW. The Capacity of the Medical Expenditure Panel Survey to Inform the Affordable Care Act. *Inquiry.* 2013;*50*(2),124–134. http://dx.doi.org/10.1177/0046958013513678

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 SB, Unangst J. Data Integration Innovations to Enhance Analytic Utility of Clinical Trial Content to Inform Health Disparities Research. *Frontiers in Oncology*. 2018;*8,365*.
http://dx.doi.org/10.3389/fonc.2018.00365

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

Hamel LM, Penner LA, Albrecht TL, Heath E., Gwede CK, Eggly S. Barriers to Clinical Trial Enrollment in Racial and Ethnic Minority Patients with Cancer. *Cancer Control*. 2016;*23*(4),327–337. http://dx.doi.org/10.1177/107327481602300404

Ludmir, E. B., Mainwaring, W., Lin, T. A., Miller, A. B., Jethanandani, A., Espinoza, A. F., Fuller, C. D. Factors associated with age disparities among cancer clinical trial participants. *JAMA Oncology*. (2019). http://dx.doi.org/10.1001/jamaoncol.2019.2055

National Academies of Sciences, Engineering, and Medicine, ed. *Communities in Action: Pathways to Health Equity.* Washington (DC): National Academies Press (US); 2017. https://www.ncbi.nlm.nih.gov/books/NBK425844/. Accessed June 11, 2020.

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

Sullivan, P.W. & Ghushchyan, V. Mapping the EQ-5D Index from the SF-12: U.S. general population preferences in a nationally representative sample
Med. Decis. Making, 26 (2006), pp. 401-409, 10.1177/0272989X06290496