Project Data Sphere Data Integration Initiatives

Steven B. Cohen and Jennifer Unangst
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. This would include connectivity to existing secondary data sources at higher levels of aggregation and via direct matches to additional health and socioeconomic measures at the individual level acquired from other sources of survey, health system, economic or administrative data. Advances in data science are also serving to facilitate the effective and efficient utilization of statistical methods in concert with big data applications to develop these enhanced analytical platforms and infrastructure. 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 would include content drawn from nonprobability based samples to enhance analytic capacity.

In this study, the content in selected Project Data Sphere LLC (PDS) cancer patient-level phase III clinical datasets have been augmented by linking the social, economic, and health-related characteristics of like cancer survivors from nationally representative health and health care-related data from the Medical Expenditure Panel Survey (MEPS). Study findings include probabilistic assessments of the representation of the patients in the respective clinical trials relative to the characteristics of cancer survivors in the general population. The study illustrates the enhancements achieved to the analytic capacity and utility of the PDS cancer clinical trial data through data integration.

Keywords: Project Data Sphere; Data Integration; Health Disparities; MEPS

1. **Introduction**

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. This would include connectivity to existing secondary data sources at higher levels of aggregation and via direct matches to additional health and socioeconomic measures at the individual level acquired from other sources of survey, health system, economic or administrative data. Advances in data science are also serving to facilitate the effective and efficient utilization of statistical methods in concert with big data applications to develop these enhanced analytical platforms and infrastructure. 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 would include content drawn from nonprobability based samples to enhance analytic capacity.

In this study, the content in selected Project Data Sphere LLC (PDS) cancer patient-level phase III clinical datasets have been augmented by linking the social, economic, and health-related characteristics of like cancer survivors from nationally representative health and health care-related data from the Medical Expenditure Panel Survey (MEPS). Study findings include probabilistic assessments of the representation of the patients in the respective clinical trials relative to the characteristics of cancer survivors in the general population. The study illustrates the enhancements achieved to the analytic capacity and utility of the PDS cancer clinical trial data through data integration.

2. Representation of Patients in Clinical Trials

Health disparities for individuals with cancer are most apparent when there are notable differences in the occurrence, frequency, death, and burden of cancer among specific population groups; these differences often are manifest when comparing the experiences of distinct racial and ethnic minority groups. While research and policy efforts have helped to reduce some observed gaps in health outcomes, cancer disparities persist (National Academies of Sciences, 2020). The driving factors for the continuance of disparities include differential access to and quality of care (NCI, 2018). Clinical trials, for example, are used to identify safe and effective treatments for all those with cancer but are often conducted among younger, healthier, and less racially diverse patients than the population at large (Hamel et al., 2016). As a result, there is an increasing interest in diversifying clinical trial patients to ensure that resultant treatments are suited for those who are disproportionately affected in the first place. As a result, there is an increasing interest in diversifying clinical trial patients to ensure that resultant treatments are suited for subgroups who are underrepresented in trials and disproportionately affected by cancer.

As noted in a recent JAMA Oncology Viewpoint article, "data sharing in clinical trials is increasingly recognized as fundamental to strengthening therapeutic research (Arfe et al, 2020)." To this end, 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 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 evaluate the efficacy of treatment-vs.-control randomizations, 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*.

These data integration efforts 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. Charter data providers include AstraZeneca, Bayer, Celgene, Janssen, Memorial Sloan Kettering Cancer Center, Pfizer, and Sanofi. 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. When additional demographic measures are available in both datasets (e.g., height, weight, body-mass index, employment status), 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.

3. Utilizing PDS and MEPS Data to Inform Health Disparities in Cancer Clinical Trials

A core component of this research effort was to determine how representative the cancer patients enrolled in clinical trials are to like cancer patients in the U.S. general population. Consequently, we focused on examining the sociodemographic and health-related characteristics of the cancer patients enrolled in specific set of phase III clinical trials relative to the characteristics of individuals in the general population with the same conditions, as represented by MEPS. We present results for one of these trials to illustrate the capacity of the PDS data when used in concert with the MEPS data.

The PDS trial utilized was D4320C00015: A Phase III, Randomised, Placebo-controlled, Double-blind Study to Assess the Efficacy and Safety of Once-daily Orally Administered ZD4054 10 mg in Non-metastatic Hormone-resistant Prostate Cancer Patients (ClinicalTrial.gov ID NCT00626548). The PDS data available for this trial represent 677 comparator patients. For comparison, prostate cancer survivors were identified among all MEPS cases from the 2000–2016 MEPS-HC Survey Full Year Consolidated Data files using the variables ICD9CODX and CCCODEX or ICD10CDX on the Medical Conditions File for the 2000-2015 or 2016 MEPS cases, respectively; it was necessary to link the Full Year Consolidated Data files with the Medical Conditions file to obtain ICD9CODX, CCCODEX, and ICD10CDX. MEPS cases with ICD9CODX = 185 or CCCODEX = 029 for 2000-2015 MEPS or ICD10CDX = C61 for 2016 MEPS were identified as prostate cancer survivors.

2,207 MEPS prostate cancer survivors were identified for representational comparisons and for linkage to the 677 PDS prostate cancer patients enrolled in the comparator arm of the trial. Because the set of prostate cancer survivors represented in the pooled MEPS data sets are representative of the prostate cancer survivors in the nation, the results of the PDS-MEPS data profiles permitted assessments of the sociodemographic and health-related characteristics that differentiated patients more likely to be represented in the trial. As noted, clinical trials are often conducted among younger, healthier, and less racially diverse patient populations than the population at large (Ludmir et al., 2019; De Moor et al., 2016; Hamel et al., 2016; O'Keefe et al., 2015). Consequently, research efforts that focus on the determinants of health disparities depend on the availability of information that distinguish cancer patients by demographic and socioeconomic factors, their access to health care services and treatments, and their health behaviors. Comparing the demographic and health characteristics of the prostate cancer patients in the comparator arm of trial NCT00626548 revealed significant departures from their representation in the nation. Prostate cancer patients in the trial were more likely to be elderly over the age of 65, Asian, Hispanic or multiracial, and less likely to be Black or White relative to the representation of prostate cancer survivors in the United States. Prostate cancer patients in the trial were significantly more likely to have better health states and also less likely to have chronic conditions such as hypertension, diabetes, asthma, arthritis or coronary heart disease relative to the profiles of prostate cancer survivors in the nation (Table 1).

4. Summary

An examination of the demographic composition of distinct sets of cancer patients in the data enhanced PDS trials revealed significant departures from their representation in the nation. Cancer patients in the trials were often more likely to be younger, white, and male in contrast to the representation of cancer survivors in the United States. Cancer patients in the PDS trials were also significantly more likely to have better health states and also less likely to have chronic conditions such as hypertension, diabetes, asthma, arthritis, or coronary heart disease relative to the profiles of cancer survivors in the nation (Cohen and Unangst, 2018).

While cancer researchers continue to advance new discoveries and treatment protocols, millions of lives continue to be lost to cancer each year. The pace of progress in improving health outcomes in cancer patients is further challenged when addressing health disparities that impact specific populations such as racial minorities and economically disadvantaged population subgroups. Health disparities for individuals with cancer are most apparent when there are notable differences in the occurrence, frequency, burden of cancer and mortality rates among specific population groups. The analytically enhanced integrated data will help researchers explore the influence of healthcare access, socioeconomic factors, and health behaviors on the patient-level representativeness and outcomes data contained in the trials included in the PDS data enclave. Researchers can now access the data and supporting documents at https://data.projectdatasphere.org/projectdatasphere/html/landing/rti . As additional clinical trial datasets are added to the PDS website, researchers can also initiate future data augmentations using MEPS by implementing the delineated linkage methodology. This project further enables researchers to use the content enriched PDS datasets to stimulate new research findings and generate insights into the representational disparities that exist in trial study designs, thus helping to improve future study designs and to help promote equity in cancer research.

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 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

de Moor, J. S., Virgo, K. S., Li, C., Chawla, N., Han, X., Blanch-Hartigan, D., et al. Access to cancer care and general medical care services among cancer survivors in the United States: An analysis of 2011 medical expenditure panel survey data. *Public Health Reports (Washington, D.C.)*, *131*(6), 783–790. (2016). http://dx.doi.org/10.1177/0033354916675852

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

Table 1. Distribution Comparison of Age, Race/Ethnicity, BMI, EQ-5D, and Health Conditions among All 2000-2016 MEPS and PDS Prostate Cancer Cases Aged 40+

	MEPS 2000–2016 Prostate Cancer Cases		**PDS Prostate Cancer Cases**			**North America PDS Prostate Cancer Cases**		
	Overall (n=2,207)		**Overall (n=677)**			**(n=127)**		
Measure	**Unweighted Count**	**% (SE)**	**Unweighted Count**	**% (SE)**	**Test of Overall MEPS v. Overall PDS Wald X^2 (p-value)**	**Unweighted Count**	**% (SE)**	**Test of North America PDS v. Overall PDS Wald X^2 (p-value)**
AGE					34.78 (pval=0.0000)			3.02 (pval=0.5542)
1 40–59	251	11.5 (1.07)	42	6.2 (0.93)		5	3.9 (1.73)	
2 60–64	251	11.7 (1.04)	65	9.6 (1.13)		13	10.2 (2.69)	
3 65–69	373	16.2 (1.18)	121	17.9 (1.47)		27	21.3 (3.63)	
4 70–74	419	19.6 (1.22)	157	23.2 (1.62)		32	25.2 (3.86)	
5 75+	913	41.0 (1.70)	292	43.1 (1.90)		50	39.4 (4.34)	
RACE/ETHNICITY					4638.83 (pval=0.0000)			448.70 (pval=0.0000)
1 WHITE	1432	80.9 (1.13)	396	58.5 (1.90)		110	86.6 (3.02)	
2 BLACK	490	12.0 (0.91)	5	0.7 (0.33)		1	0.8 (0.78)	
3 HISPANIC	201	4.4 (0.53)	48	7.1 (0.99)		4	3.1 (1.55)	
4 ASIAN/PACIF	54	1.6 (0.38)	167	24.7 (1.66)		2	1.6 (1.11)	
5 OTHER/MULTI	30	1.1 (0.32)	61	9.0 (1.10)		10	7.9 (2.39)	
BMI					9.29 (pval=0.0257)			25.39 (pval=0.0000)
1– UNDERWEIGHT	19	0.7 (0.21)	3	0.4 (0.26)		1	0.8 (0.79)	
2– NORMAL	537	25.5 (1.41)	195	28.9 (1.75)		18	14.3 (3.12)	
3– OVERWEIGHT	935	45.5 (1.54)	309	45.8 (1.92)		58	46.0 (4.44)	
4– OBESE	604	28.3 (1.54)	168	24.9 (1.67)		49	38.9 (4.35)	
EQ5D Score Decile Categories					69.38 (pval=0.0000)			4.22 (pval=0.5188)
1 <= 0.656	327	14.6 (1.12)	76	11.8 (1.27)		16	13.0 (3.04)	
2 >0.656 and <= 0.725	228	10.4 (0.80)	83	12.8 (1.32)		17	13.8 (3.11)	
3 >0.725 and <= 0.779	210	9.7 (0.79)	44	6.8 (0.99)		11	8.9 (2.57)	
4 >0.779 and <= 0.814	170	9.1 (0.83)	87	13.5 (1.34)		22	17.9 (3.46)	
5 >0.814 and <= 0.883	304	15.6 (1.02)	87	13.5 (1.34)		14	11.4 (2.87)	
6 >0.883	776	40.6 (1.56)	269	41.6 (1.94)		43	35.0 (4.30)	
HIGH BLOOD PRESSURE					808.09 (pval=0.0000)			4.54 (pval=0.0330)
1 Yes	1494	67.9 (1.49)	172	25.4 (1.67)		23	18.1 (3.42)	
2 No	667	32.1 (1.49)	505	74.6 (1.67)		104	81.9 (3.42)	
CORONARY HEART DISEASE					195.65 (pval=0.0000)			0.39 (pval=0.5344)
1 Yes	469	22.7 (1.40)	22	3.2 (0.68)		3	2.4 (1.35)	
2 No	1690	77.3 (1.40)	655	96.8 (0.68)		124	97.6 (1.35)	
DIABETES					94.47 (pval=0.0000)			0.99 (pval=0.3209)
1 Yes	500	19.4 (1.43)	37	5.5 (0.87)		10	7.9 (2.39)	
2 No	1660	80.6 (1.43)	640	94.5 (0.87)		117	92.1 (2.39)	
ASTHMA					49.49 (pval=0.0000)			0.37 (pval=0.5416)
1 Yes	166	6.8 (0.84)	6	0.9 (0.36)		2	1.6 (1.11)	
2 No	1995	93.2 (0.84)	671	99.1 (0.36)		125	98.4 (1.11)	
ARTHRITIS					971.01 (pval=0.0000)			1.42 (pval=0.2340)
1 Yes	1086	51.6 (1.56)	21	3.1 (0.67)		7	5.5 (2.03)	
2 No	991	48.4 (1.56)	656	96.9 (0.67)		120	94.5 (2.03)	

MEPS = Medical Expenditure Panel Survey; SE = standard error.

MEPS prostate cancer survivors were identified among all MEPS cases from the 2000–2016 MEPS-HC Survey Full Year Consolidated Data files using the variables ICD9CODX and CCCODEX or ICD10CDX on the Medical Conditions File for the 2000–2015 or 2016 MEPS cases, respectively; it was necessary to link the Full Year Consolidated Data files with the Medical Conditions file to obtain ICD9CODX, CCCODEX, and ICD10CDX. MEPS cases with ICD9CODX = 185 or CCCODEX = 029 for 2000–2015 MEPS or ICD10CDX = C61 for 2016 MEPS were identified as prostate cancer survivors.

The MEPS full year weight PERWTXXF was divided by 17, the number of years pooled together that contained at least one prostate cancer survivor, to produce weighted estimates and SEs via SUDAAN. PDS estimates and SEs are unweighted.

For categorical measures, tests of overall MEPS or North America PDS versus overall PDS present the overall Wald chi-squared statistic for a goodness of fit test, using the PDS overall distribution as the theoretical distribution for the measure.

For estimates, N/A indicates that all values for the variable were missing. For test statistics, N/A indicates that no degrees of freedom were available due to inadequate number of observed levels.