Integration of Clinical and National Health Care Survey Data to Inform Disparities

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 health care surveys are enhanced through integrated designs that link additional medical, behavioral, environmental, socio-economic and financial content from multiple sectors. In this study, the content in selected Project Data Sphere (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. Attention is given to the identification of the segment of the target population represented by the nonprobability based clinical trial samples and restricting inferences based on the integrated data to these subdomains. 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 and an evaluation of the reproducibility of analytic findings. The study illustrates the enhancements achieved to the analytic capacity and utility of the PDS cancer clinical trial data through data integration.

Keywords: Data integration; MEPS; clinical trials; Project Data Sphere; Health Disparities

1. Introduction

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 disparities often manifest when comparing the experiences of distinct racial and ethnic minority groups. The factors that influence such differentials in patient outcomes include poverty, lack of access to prevention/detection services, and whether high-quality treatment is available. Consequently, research efforts that focus on the determinants of health disparities depend on the ability to distinguish cancer patients by demographic and socioeconomic factors, their access to health care services and treatments, and their health behaviors.

The quality and content of clinical trial data can be enhanced through integrated designs that link the data with additional medical, behavioral, environmental, socio-economic and financial content from multiple sectors. In this study, we discuss an example of one such integrated design, where we have augmented the content in selected Project Data Sphere (PDS) cancer patient-level phase III clinical datasets 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). We focus our attention on identifying the segment of the target population represented by the nonprobability based clinical trial samples and restricting inferences based on the integrated data to this subdomain. 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 and an evaluation of the reproducibility of analytic findings. The study illustrates how data integration can enhance the analytic capacity and utility of the PDS cancer clinical trial data. This research effort was made possible through funding provided by a grant from the Robert Wood Johnson Foundation.

2. Integration of Clinical and National Health Care Survey Data to Inform Health Disparities Research

This work was facilitated by integrating two rich data sources: clinical trial datasets from the Project Data Sphere online platform and public use survey data from the Medical Expenditure Panel Survey. We describe the content and design of each individual data source in more detail below.

Clinical Trial Data: Project Data Sphere, LLC (PDS) was formed in 2012 to catalyze cancer research by bringing together diverse minds and technology to help unleash the full potential of existing clinical trial data. PDS, an independent initiative of the CEO Roundtable on Cancer's (CEORT's) Life Sciences Consortium, operates a first-of-its-kind research platform that provides the research community with broad access to both de-identified patient level data from oncology clinical trials and freely available analytic tools to assist them in analyzing those data. A primary goal of PDS is to advance new research efforts that will improve the lives of cancer patients and their families around the world [1,9,11,12]. These data are rich in terms of measures that characterize the clinical trials under study, treatment protocols, and patient outcomes. However, to address confidentiality provisions inherent to the trials, data providers are required to de-identify patient-level data prior to uploading datasets to the PDS online service by masking or removing certain demographic data. Consequently, it is not currently possible to assess 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 inclusion of these measures would significantly enhance the analytic capacity and utility of the PDS data, further stimulating hypothesis generation and the initiation of new studies that explore these relationships.

ProjectDataSphere.org hosts over 200 phase III oncology clinical trial datasets, representing more than 160,000 cancer patients. This data sharing initiative has already demonstrated its benefit to the research community with triple the usage of other major, clinical trial data-sharing efforts combined. PDS data have also been cited by 21 peer-accepted publications on 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 [11,12]. For selected PDS datasets, our project has extended the utility of these publicly available data by joining PDS patient-level data with nationally representative health-related data from MEPS. MEPS is the nation's primary source of nationally representative, comprehensive, person-level data on health care use, insurance coverage, and expenses. With this additional content, the PDS data platform will further serve to advance cancer research by permitting more granular subgroup analyses and meta-analyses of related treatment protocols. This is particularly important because clinical trials are often conducted among younger, healthier, and less racially diverse patient populations than the population at large [6,7]. The augmented datasets have already enabled researchers to evaluate the efficacy of treatment-vs.-control randomizations and to investigate whether the added variables are related to outcomes of interest. Researchers can also conduct probabilistic assessments of the proportion of the U.S. population that the cancer patient outcomes observed in the PDS online service may or may not represent.

Survey Data: MEPS is characterized by an integrated survey design. Since its inception, the primary analytical focus of MEPS has been health care access, coverage, cost, and use. Over the past several years, MEPS data have supported a highly visible set of descriptive and behavioral analyses of the U.S. health care system. These include studies of the population's access to, use of, and expenditures and sources of payment for health care; the availability and costs of private health insurance in the employment-related and non-group markets; the population enrolled in public health insurance coverage versus those without health care coverage; and the role of health status in health care use, expenditures, household decision making, and in health insurance and employment choices. Because of the breadth of MEPS data, these data have informed the nation's economic models and their projections of health care expenditures and utilization. The level of cost and coverage detail in MEPS data has enabled public and private sector economic models to develop national and regional estimates of the impact of changes in financing, coverage, and reimbursement policy, as well as estimates of who benefits and who bears the cost of a change in policy. MEPS has been collecting data on health care utilization and expenditures annually since 1996. The survey is sponsored by the Agency for Healthcare Research and Quality. In addition to collecting nationally representative data to yield annual estimates for a variety of measures related to health care use and expenditures, MEPS provides estimates related to health status, demographic characteristics, employment, health insurance coverage, and access to health care. 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. 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. Respondents identify medical providers from whom they have received services [2-5].

The set of households selected for MEPS-HC is a subsample of approximately 15,000 households/35,000 individuals participating in the National Health Interview Survey (NHIS). NHIS is an ongoing annual household survey of ~40,000 households conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention, to obtain national estimates of health care utilization, health conditions, health status, insurance coverage, and access representing the civilian noninstitutionalized population. In addition to the cost savings achieved by eliminating the need to independently list and screen households, selecting a subsample of NHIS participants has led to enhanced analytical capacity of the resultant survey data. Use of NHIS data in concert with the data collected for MEPS provides greater capacity for longitudinal analyses not otherwise available. Furthermore, the large number and dispersion of the primary sampling units in MEPS has resulted in more precise expenditure survey designs. The MEPS-HC survey consists of an overlapping panel design in which any given sample panel is interviewed a total of five times in person 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.

Methods: To create the enhanced dataset referenced specially in this paper, patients from the PDS clinical trial dataset *LungNo_MerckKG_2007_145* were linked with similar lung cancer survivors from the MEPS 2000-2013 survey data. Because we present a detailed summary of the data integration methodology in a related publication [2], we do not discuss the linkage approach here. Using MEPS 2000-2013 data, we identified 653 lung cancer survivors as candidates for linkage to the 507 PDS lung cancer patients enrolled in the comparator arm of the Merck trial. Using the hierarchical linkage methodology described in our related publication, 401 of the 507 PDS lung cancer patients obtained a linkage with at least one lung cancer survivor represented in the MEPS. Alternatively, 401 of the 653 lung cancer survivors in MEPS achieved at least one linkage to PDS lung cancer cases. This observed differential in linkage rates, conditioned on the characteristics of the cancer patients in the respective datasets, was suggestive of the distinct patient selection criteria that distinguish these trials.

3. Characteristics of Lung Cancer Cases in PDS Clinical Trials vs. Cancer Survivors in the Population

Once the PDS cancer survivor data were linked with national health care data from MEPS, the analytical aims of the study could be addressed. A core component of this research effort was to determine how representative the cancer patients enrolled in clinical trials were to like cancer patients in the general population. Consequently, we focused on examining the sociodemographic and health-related characteristics of those cancer patients enrolled in specific phase III clinical trials relative to the characteristics of individuals in the general population with the same conditions.

We illustrate the analytic capacity of the enhanced datasets using comparator arm patients from a PDS lung cancer clinical trial: *LungNo_MerckKG_2007_145*. Because the set of lung cancer survivors represented in the pooled MEPS data sets are representative of the lung cancer survivors in the nation, the results of the PDS-MEPS data linkage permitted assessments of the sociodemographic and health-related characteristics that differentiated patients more likely to be represented in the trial. For these analyses, a logistic regression model was specified to determine the most salient factors that differentiated patients in the PDS trial from their lung cancer survivor counterparts in the overall population. More specifically, lung cancer survivors represented in MEPS who linked with the patients in the PDS trial were classified as Y = 1, and the unlinked cancer survivors in MEPS were classified as Y = 0. The following sociodemographic and health-related measures were included in the model to determine their significance in distinguishing the likelihood of representation in the PDS lung cancer trial under study (Table 1):

- *Sociodemographic:* Age, race/ethnicity, sex, marital status, employment status, education level, income level, year in MEPS
- *Access related:* Health insurance coverage, ability to obtain necessary medical care
- *Health related:* EQ-5D quality of life index score, perceived health status, limitations in physical functioning, smoker status
- *Health care related:* Office-based physician visits, in-patient hospital stays, emergency room visits, prescription drug purchases, total health care expenditures

Table 1: Measures considered as potential predictors of trial linkage status for MEPS lung cancer survivors

Measures	Description
Age	Age in years at end of the MEPS survey year
Race	White, Black, Other (including Hispanic)
Sex	Male, Female
EQ-5D Decile Category	For MEPS 2000–2003, the categorized predicted value of EQ-5D based on Dolan prediction equation. For MEPS 2004–2013, the categorized predicted value of EQ-5D based on Sullivan-Ghushchyan prediction model. Fewer than ten decile categories resulted due to ties. $-0.016 \leq$ EQ-5D < 0.620 $0.620 \leq$ EQ-5D < 0.689 $0.689 \leq$ EQ-5D < 0.725 $0.725 \leq$ EQ-5D < 0.760 $0.760 \leq$ EQ-5D < 0.796 $0.796 \leq$ EQ-5D < 0.848 $0.848 \leq$ EQ-5D < 0.883 $0.883 \leq$ EQ-5D < 1.000 EQ-5D ≥ 1.000
Marital Status	Married, Not married (including divorced, separated, widowed, never married)
Employment Status	Not employed, Employed at any time during reference period
Education Level	No degree, Earned at least GED or high school diploma
Income Level	High income (family income ≥ 400% of the poverty level), poor through middle income (family income < 400% of the poverty level)
MEPS Survey Period	2000–2003, 2004–2013
Health Insurance Coverage	Any private insurance, Public insurance only, Uninsured
Smoker Status	Current smoker, Not current smoker
Perceived Health Status	Excellent/Very Good/Good, Fair/Poor
Limitation in Physical Functioning	Yes, No
Number of Prescribed Medicine Purchases	Frequency in year
Number of Hospital Discharges	Frequency in year
Number of Emergency Room Visits	Frequency in year
Number of Office-based Physician Visits	Frequency in year
Total Health Care Expenditures	Continuous measure for year
Access to Necessary Medical Care	Able to get access, Unable to get access

Source: Medical Expenditure Panel Survey Household Component Data Files 2000–2013, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

Our multivariate logistic regression analyses helped identify the set of significant factors ($p < 0.05$) that were more characteristic of the PDS lung cancer patients enrolled in the trial relative to adult lung cancer survivors in the U.S. noninstitutionalized population (Table 2). Based on the results of the logistic model, the following measures were identified as significant predictors ($p < 0.05$) 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.

Table 2. Logistic regression model to identify factors associated with trial linkage status for MEPS lung cancer survivors

Independent Variables and Effects	Beta Coeff.	SE Beta	*p*-value t-test B = 0	d.f.	Wald F	*p*-value Wald F
Overall Model				9	6.95	< 0.0001
Intercept	1.06	0.43	0.0145			
Marital Status				1	6.16	0.0134
Married	1.01	0.40	0.0134			
Sex				1	11.04	0.0010
Female	−1.61	0.49	0.0010			
MEPS Survey Year				1	6.47	0.0113
2000-2003	−0.93	0.37	0.0113			
EQ-5D Decile Category	0.26	0.07	0.0001	1	14.81	0.0001
Race				2	25.94	< 0.0001
Other (including Hispanic)	-4.39	0.85	< 0.0001			
Black	-5.93	0.91	< 0.0001			
Access to Necessary Medical Care				1	3.17	0.0758
Unable to get access	1.09	0.61	0.0758			
Smoking Status				1	4.46	0.0352
Current smoker	0.94	0.44	0.0352			

Notes: *n* = 470
Analysis performed using SUDAAN statistical software. The s*ubpopn* statement was used to conduct the subpopulation analysis of lung cancer survivors from among all MEPS cases (2000–2013).
Pseudo R-square: 0.429977
−2 * Normalized log-likelihood with intercepts only: 580.32
−2 * Normalized log-likelihood full model: 316.14
Approximate chi-square (−2 * log-L ratio): 264.18
Degrees of freedom: 8
Denominator degrees of freedom: 445
Source: Medical Expenditure Panel Survey Household Component Data Files 2000–2013, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

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. In these analyses, the standard errors of the survey estimates and model coefficients derived from MEPS have been adjusted for the impact of clustering due to the multistage survey design, and the test statistics used to test for equivalence in estimates and significance in model coefficients have also been adjusted to control for survey design complexities.

4. Summary

This project enhanced the data profiles in selected patient-level cancer phase III clinical datasets hosted on the PDS online service by linking the social, economic, and health-related characteristics of cancer survivors from the MEPS, a nationally representative health and health care–related survey. With this additional content, the PDS data platform further serves to advance cancer research initiatives that permit more granular subgroup and meta-analyses of related treatment protocols. Clinical trials are often conducted among younger, healthier, and less racially diverse patient populations than the population at large. Potential analyses with this analytically enhanced PDS database include probabilistic assessments of the proportion of the population in the nation that the cancer patient outcomes observed in the PDS online service may or may not represent.

Using data integration methods, this study linked sociodemographic, access, health, and health care-related measures associated with a nationally representative set of lung cancer survivors included in MEPS to similar cancer patients in the PDS analytic datasets. In addition to utilizing demographic information (age, race/ethnicity, and sex) available in both data sources, the data integration was further advanced by including responses to patient-reported outcomes data captured in the EQ-5D index score derived from the EuroQoL five-dimensions questionnaire. The data integration with MEPS now facilitates the inclusion of content on demographic characteristics (education level, marital status, family structure); socioeconomic measures (income, poverty status); and health and health care-related measures (health status, number of chronic conditions, access to care, health insurance, medical utilization, and expenditures).

The measures appended to each patient-level record in selected datasets hosted on PDS are in the form of data vectors or distributions derived from MEPS, including a weight that supports population-level inference. Comparison of these data vectors and these families of distributions will help enable researchers to investigate whether the added measures potentially affect cancer patient outcomes.

Note: This research effort was made possible as a consequence of funding provided by a grant from the Robert Wood Johnson Foundation. This effort reflects a collaboration between RTI International, Project Data Sphere and the Robert Wood Johnson Foundation. Special acknowledgements go to Dave Handelsman, Project Data Sphere for his contributions.

References

Abdallah, K., C. Hugh-Jones, T. Norman, S. Friend and G. Stolovitzky. 2015. The Prostate Cancer DREAM Challenge: A Community-Wide Effort to Use Open Clinical Trial Data for the Quantitative Prediction of Outcomes in Metastatic Prostate Cancer. Oncologist. 20(5):459-60.

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*. doi: 10.3389/fonc.2018.00365

Cohen, S. B. & J. Cohen, 2013. "The Capacity of the Medical Expenditure Panel Survey to Inform the Affordable Care Act", Inquiry. 50(2):124-34

Cohen, J., S. Cohen, and J. Banthin. 2009. "The Medical Expenditure Panel Survey: A National Information Resource to Support Healthcare Cost Research and Inform Policy and Practice." Medical Care 47 (7, Suppl. 1): 44–50.

Cohen, S., and T. Buchmueller. 2006. "Trends in Medical Care Costs, Coverage, Use and Access: Research Findings from the Medical Expenditure Panel Survey." Medical Care 44 (5): 1–3.

De Moor JS, Virgo KS, 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 Rep*. (2016) 131:783–90. doi: 10.1177/0033354916675852

DeSantis CE, Siegel RL, Sauer AG, Miller KD, Fedewa SA, Alcaraz KI, et al. Cancer statistics for African Americans, 2016: progress and opportunities in reducing racial disparities. *CA Cancer J Clin*. (2016) 66:290–308. doi: 10.3322/caac.21340

Greene, A., K. Reeder-Hayes, R. Corty, E. Basch, M. Milowsky, S. Dusetzina, A. Bennett and W. Wood. 2015. “The Project Data Sphere Initiative: Accelerating Cancer Research by Sharing Data.” The Oncologist 20 (5): 464-e20.

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) 23:327–37. doi: 10.1177/107327481602300404

O'Keefe EB, Meltzer JP, Bethea TN. Health disparities and cancer: racial disparities in cancer mortality in the United States, 2000-2010. *Front Public Health* (2015) 3:51. doi: 10.3389/fpubh.2015.00051

Project Data Sphere. Available online at: https://projectdatasphere.org/projectdatasphere/html/home (Accessed March 8, 2018).

Project Data Sphere. *Current Project Data Sphere List of Peer-Accepted Publications* (2017). Available online at: https://projectdatasphere.org/projectdatasphere/html/WhatsNewPress

White-Means SI, Osmani AR. Racial and ethnic disparities in patient-provider communication with breast cancer patients: evidence from 2011 MEPS and experiences with cancer supplement. *Inquiry* (2017) 54:46958017727104. doi: 10.1177/0046958017727104

Yabroff K. R., E. Dowling, J. Rodriguez, D. Ekwueme, H. Meissner, A. Soni, G. Lerro, G. Willis, L. Forsythe, L. Borowski and K. Virgo. 2012. “The Medical Expenditure Panel Survey (MEPS) experiences with cancer survivorship supplement.” Journal of Cancer Survivorship. 6(4):407-19