Data Integration Enhancements to Project Data Sphere's Analytic Capacity

Steven B. Cohen and Jennifer Unangst

RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Project Data Sphere (PDS) is a research platform that provides the research community with broad access to both de-identified patient-level data from oncology clinical trials and related analytic tools. While these data are rich in terms of measures that characterize the clinical trials under study, data providers are required to de-identify patient-level data by removing key demographic data. To address these analytic constraints, the data profiles in selected PDS patient-level cancer 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 healthcare-related survey data. Using statistical matching and model-based techniques, patient-level records in selected PDS datasets have been linked to comparable cancer survivors, and thereby augmented with survey content on social, economic and health related characteristics. This article will provide an overview of the methodologies used to join PDS patient-level data with nationally representative health-related data on cancer survivors from the MEPS and an evaluation of the stability of analytic results.

Keywords: Project Data Sphere, data integration, MEPS, clinical trials

1. Introduction

Project Data Sphere (PDS) is a research platform that provides the research community with broad access to both de-identified patient-level data from oncology clinical trials and related analytic tools. While these data are rich in terms of measures that characterize the clinical trials under study, data providers are required to de-identify patient-level data by removing key demographic data. To address these analytic constraints, the data profiles in selected PDS patient-level cancer 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 healthcare-related survey data. Using statistical matching and model-based techniques, patient-level records in selected PDS datasets have been linked to comparable cancer survivors, and thereby augmented with survey content on social, economic and health related characteristics. This article will provide an overview of the methodologies used to join PDS patient-level data with nationally representative health-related data on cancer survivors from the MEPS and an evaluation of the stability of analytic results.

2. Analytical Enhancements Achieved Through Linkage of Surveys to Other Sources of Data

Cancer researchers continue to advance new discoveries and treatment protocols, yet every year, millions of lives are lost to cancer. With researchers working independently and with

declining resources, solutions are not advancing quickly enough. 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]. 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, 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 cannot currently be assessed. 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.

Our primary goal is to create a collection of enhanced research databases that will add significant socioeconomic and health care access content to the existing datasets hosted on the PDS online service, thereby enhancing their analytic capacity and utility. This data enhancement project will serve to further advance the mission of the PDS platform by enabling new explorations into the potential influence of health care access, socioeconomic factors, and health behaviors on the patient-level efficacy and outcomes data contained in the PDS online service. This data integration effort will generate collective insights that may yield improvements in trial designs and stimulate new research findings derived from applying advanced analytic methodologies to the content-enhanced datasets. This research effort was made possible as a consequence of funding provided by a grant from the Robert Wood Johnson Foundation.

Objectives and Activities: The data profiles in selected patient-level cancer phase III clinical datasets hosted on the PDS online service are being augmented by linking the social, economic, and health-related characteristics of cancer survivors from nationally representative health and health care–related survey data. Through the application of statistical linkage and model-based techniques, patient-level records in selected PDS datasets are being linked to comparable cancer survivors and thereby augmented with survey content on social, economic, and health-related characteristics. Specifically, we are joining PDS patient-level data with nationally representative health-related data from the Medical Expenditure Panel Survey (MEPS), 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 would further serve 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. The augmented datasets should enable researchers to evaluate the efficacy of treatment-vs.-control randomizations and to investigate whether the added variables are related to outcomes of interest. Other potential impacts 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. The data in the PDS enclave cannot currently support these types of investigations. The addition of the MEPS data to the patient-level data within

the PDS enclave will facilitate hypothesis-generating research efforts that explore the level of variation in patient outcomes potentially attributable to differentials in access to basic health care services and their utilization, to socioeconomic characteristics, and to health behaviors and preferences. It will support exploratory analyses designed to examine questions such as How are variations in cancer patients' access to health care and income impacting patient outcomes in specific phase III clinical trials? What variations in patient outcomes are associated with specific demographic, socioeconomic, and health-related factors? Are the demographic characteristics of those cancer patients enrolled in specific phase III clinical trials comparable to cancer patients with the same disease in the general population?

3. Applications to the Medical Expenditure Panel Survey

One of the core health care surveys in the United States, the Medical Expenditure Panel Survey (MEPS), is characterized by a consolidated survey design. Since its inception, the primary analytical focus of the MEPS has been directed to the topics of health care access, coverage, cost and use. 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. 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 and those without health care coverage; and the role of health status in health care use, expenditures, and household decision making, and in health insurance and employment choices. As a consequence of its breadth, the data have informed the nation's economic models and their projections of health care expenditures and utilization. The level of the cost and coverage detail collected in the MEPS 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.

The Medical Expenditure Panel Survey (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 (AHRQ). In addition to collecting nationally representative data to yield annual estimates for a variety of measures related to health care use and expenditures, the MEPS also provides estimates related to health status, demographic characteristics, employment, health insurance coverage, and access to health care. The MEPS consists of a family of three interrelated surveys: The Household Component (MEPS-HC), the Medical Provider Component (MEPS-MPC), and the Insurance Component (MEPS-IC). The MEPS-IC also collects establishment-level data on insurance programs. Through a series of interviews with household respondents, the 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 [4-6, 13].

The set of households selected for the Household Component is a subsample of those participating in the National Health Interview Survey (NHIS), an ongoing annual household survey of approximately 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. In addition to the cost savings achieved by eliminating the need to independently list and screen households, selecting a subsample of NHIS participants has resulted in an enhancement in analytical capacity of the resultant survey data. Use of the NHIS data in concert with the data collected for the MEPS provides an additional capacity for longitudinal analyses not otherwise available. Furthermore, the large number and dispersion of the primary sampling units in MEPS has resulted in improvements in precision over prior expenditure survey designs. The MEPS HC survey consists of an overlapping panel design in which any given sample panel is interviewed a total of 5 times in person over 30 months to yield annual use and expenditure data for two calendar years. These rounds of interviewing are spaced about 5 to 6 months apart. The interview is administered through a computer assisted personal interview mode of data collection, and takes 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 Medical Provider Component is a survey of the medical providers, facilities and pharmacies that provided care or services to sample persons. The primary objective is to collect detailed data on the expenditures and sources of payment for the medical services provided to individuals sampled for the MEPS. Such data are essential to improve the accuracy of the national medical expenditure estimates derived from the MEPS, since household respondents are not always the most reliable source of information on medical expenditures. MPC data are collected a year after the household health care event information is collected to allow adequate time for billing transactions to be completed. The MPC collects data on dates of visits/services, use of medical care services, charges, sources of payments and amounts, and diagnoses and procedure codes for medical visits/encounters. Only providers for whom a signed permission form was obtained from the household authorizing contact are eligible for data collection in the MPC. The categories of providers in the MPC include (1) office-based medical doctors; (2) hospital facilities providing inpatient, outpatient, and emergency room care; (3) health maintenance organizations (HMOs); (4) physicians providing care during a hospitalization; (5) home care agencies; and (6) pharmacies. RTI International is the data collection organization for the MEPS MPC.

In 2016, a linked Medical Organization Survey (MEPS-MOS) was added to the MEPS. The principal objectives of this MEPS design enhancement were (1) to develop procedures for identifying the medical organizations associated with the usual source of office-based ambulatory care physicians from whom a nationally representative sample of individuals receive medical care; (2) to refine a survey questionnaire designed for assessing important features of the staffing, organization, policies, and financing of office-based and related ambulatory care medical care providers; (3) to collect organizational level data associated with these providers of medical care to MEPS respondents; (4) to develop estimation weights that support nationally representative linked provider-respondent data based on the MEPS-MOS survey; and (5) to make the linked provider-respondent data set available to the research community.

3.1 Research Method:

The core datasets that are being used for this project consist of historical, patient-level data from academic and industry phase III cancer clinical trials available in the PDS online service and public use files from MEPS. All project members of the team have approved access to the phase III cancer clinical trial data. The MEPS data files are accessible for downloading at the MEPS website:

https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp.

Furthermore, as noted from the 2013 MEPS public use file, which is comparable to other existing MEPS annual sample sizes, there are more than 2,000 sample adults aged 18 and older with a reported cancer diagnosis available for statistical linkage in each year. In addition, there are more than 225 sample adults with a reported prostate cancer diagnosis, more than 120 sample adults with a reported colon cancer diagnosis, more than 330 sample adults with a reported breast cancer diagnosis, and more than 130 sample adults with a cervical cancer diagnosis.

The planned statistical linkage between the MEPS and PDS data will utilize variables available in both datasets. In addition to demographic data on cancer patients' age, race, and sex, several of the datasets hosted on PDS include EQ-5D™. The EQ-5D™ descriptive system consists of the following five health-related components: Mobility, Self-care, Usual activities, Pain/discomfort, and Anxiety/depression. Each dimension has three levels, reflecting no health problems, moderate health problems, and extreme health problems. A measure for which there are no problems has a level 1 specification, while a component for which there are extreme problems has a level 3 response. Consequently, there are $3^5 = 243$ health states defined by the instrument, with the associated 5-digit response profiles ranging from 11111 for perfect health to 33333 for the worst possible state. To calculate the EQ-5D™ index score based on the U.S. population-based preference weights, a scoring algorithm has been created and operationalized. For the U.S. general population, the possible EQ-5D™ index scores range from -0.11 (i.e., 33333) to 1.0 (i.e., 11111) on a scale where 0.0 = death and 1.0 = perfect health [9]. The EQ-5D has also been administered in the past in the MEPS, which also includes administration of the 12-Item Short Form Health Survey (SF-12) developed from the Rand Medical Outcomes Study. The SF-12 is a general health status instrument with 12 questions producing two summary scores, the Physical Component Summary (PCS-12) and the Mental Component Summary (MCS-12). These scores are determined for each adult sample participant in MEPS, based on their responses to the SF-12. These respective components are scored such that higher scores represent better physical and emotional function and are standardized whereby the mean score is 50 and standard deviation is 10 in the general population. Using MEPS responses from the SF-12, predicted values of the EQ-5D index scores can be derived from MEPS using an algorithm developed by Sullivan and Ghushchyan (2006) that only requires the availability of the MCS-12 and PCS-12 scores [2, 11]. Thus, the statistical linkage will use a set of discriminatory variables that includes age, race, and sex, and the predicted values of the EQ-5D index scores. When additional demographic measures are available in the PDS for this statistical linkage (e.g., height, weight, BMI, employment status), they will also be incorporated in the process. Several years of MEPS data on cancer survivors could be pooled to enhance the sample sizes of cases available for linkage for specific cancer classifications. Options for linkage will permit 1-1, many-1, and many-many aggregations. Particular attention is being given to ensuring that the confidentiality provisions of both data sources are satisfied. Several approaches are being considered to implement the statistical linkage between the MEPS and select PDS datasets that cover the more prevalent cancers [8, 12].

4. Example of Linkage of PDS Lung Cancer Patients and MEPS Data

PDS data file *LungNo_MerckKG_2007_145* includes 507 lung cancer patients, representing the intent to treat population. Age, sex, race, and measures of the EQ-5D were

used to link to MEPS cases. Each PDS patient completed the EQ-5D questionnaire at multiple points during the study (e.g., at screening, during treatment, at end of study, and possibly multiple times during posttreatment phase), so it was necessary to assign a single health state to each patient prior to linking with the MEPS data. The five dimensions of EQ-5D at baseline were used to derive the EQ-5D summary scores for linkage. Baseline measurements were identified using QSGRPID = "EQ5D - WEEK0."

MEPS lung cancer survivors were identified among all MEPS cases from the 2000-2013 Household Component (HC) Survey Full Year Consolidated Data files using the variable ICD9CODX on the Medical Conditions File; it was necessary to link the Full Year Consolidated Data files with the Medical Conditions file to obtain ICD9CODX. MEPS cases with ICD9CODX = 162 were identified as lung cancer survivors.

MEPS lung cancer cases with a non-positive person-level weight (PERWTF) were ineligible for inclusion in the linkage process and are not represented in the linked dataset. Table 1 shows the number of MEPS lung cancer cases deemed eligible for linkage; this represents the set of MEPS cases included in the linked dataset. Since MEPS is a panel survey, it is possible that an individual may be represented in multiple years (maximum of two years).

Table 1. Number of MEPS Lung Cancer Survivors Eligible for Linkage by MEPS Year

Year	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2000	28	4.29	28	4.29
2001	37	5.67	65	9.95
2002	49	7.50	114	17.46
2003	46	7.04	160	24.50
2004	46	7.04	206	31.55
2005	36	5.51	242	37.06
2006	33	5.05	275	42.11
2007	49	7.50	324	49.62
2008	60	9.19	384	58.81
2009	53	8.12	437	66.92
2010	61	9.34	498	76.26
2011	59	9.04	557	85.30
2012	49	7.50	606	92.80
2013	47	7.20	653	100.00

Age, sex, race, and measures of the EQ-5D were used to link to PDS cases.

4.1 EQ-5D Estimation Methods

For PDS and MEPS 2000-2003, the five dimension measures (mobility, self-care, anxiety/depression, pain/discomfort, and usual activities) of the EQ-5D were available. Thus, it was possible to directly score a summary value of the EQ-5D (EQ5DDIRECT) using an algorithm developed by [9]. Additionally, the five measures were used to obtain a predicted value of the EQ-5D (EQ5DDOLAN) based on a modeling approach developed by Dolan [3]. For MEPS 2000-2003, the predicted EQ-5D value from the Dolan model was already provided on the source MEPS data files (EQU42). This value was validated,

so both the original value from the MEPS data files (EQU42) and the recalculated value from validation (EQ5DDOLAN) are available on the linked dataset.

For MEPS 2004-2013, only the Physical and Mental Component Summary scores (PCS42, MCS42) from the MEPS Short Form-12 Questionnaire on health status and health care quality were available to calculate a predicted EQ-5D summary score. This prediction method is based on a modeling approach developed by Sullivan and Ghushchyan [2].

A sequential hierarchical approach was used to link PDS cases to MEPS cases. Each step of the approach represents some degree of relaxation for the linkage criteria, such that linkages obtained at an earlier step are stricter than those obtained at a later step. A distinct approach was used for MEPS 2000-2003 versus MEPS 2004-2013, since the available EQ-5D summaries differed between these sets.

To link PDS cases with MEPS 2000-2003, a three-step approach was used.

- The first step required exact matches on single year age, sex, race, and the EQ-5D value directly scored from the five measures.
- The second step required exact matches on categorized age, sex, race, and the EQ-5D value directly scored from the five measures. Age categories included 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, and 85+.
- The third step required exact matches on collapsed categorized age, sex, race, and the decile categories of the predicted EQ-5D values. Collapsed age categories included 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+.

To link PDS cases with MEPS 2004-2013, a two-step approach was used.

- The first step required exact matches on single year age, sex, race, and the decile categories of the predicted EQ-5D values.
- The second step required exact matches on collapsed categorized age, sex, race, and the decile categories of the predicted EQ-5D values. Collapsed age categories included 18-24, 25-34, 35-44, 45-54, 55-64, 65-74, and 75+.

5. Summary

This project enhances 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. The data in the PDS enclave cannot currently support these types of data due to confidentiality constraints. With this additional content, the PDS data platform would further serve 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.

The integrated PDS -MEPS data should also facilitate exploratory analyses designed to examine questions such as

- How are variations in cancer patients' access to health care and income impacting patient outcomes in specific phase III clinical trials?

- What variations in patient outcomes are associated with specific demographic, socioeconomic, and health-related factors?

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 and Alan Karr, RTI International for their contributions.

References

[1] 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.

[2] AHRQ. *Calculating the U.S. Population-based EQ-5D™ Index Score.*" 2005. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/rice/EQ5Dscore.htm

[3] Dolan, P. "*Modeling Valuations for EuroQol Health States."* 1997. Medical Care, Vol. 35, No. 11, pp. 1095-1108. http://pauldolan.co.uk/wp-content/uploads/2011/07/modelling-valauations-for-health.pdf

[4] 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

[5] 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.

[6] 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.

[7] Dolan, P. "*Modeling Valuations for EuroQol Health States."* 1997. Medical Care, Vol. 35, No. 11, pp. 1095-1108. http://pauldolan.co.uk/wp-content/uploads/2011/07/modelling-valauations-for-health.pdf

[8] Fellegi, I. P., and A. Sunter. 1969. "A theory for record linkage". Journal of the American Statistical Association, 64, 1183-1210.

[9] 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.

[10] Shaw J. W., J. Johnson and S. Coons. 2005. "U.S. valuation of the EQ-5D™ health states: development and testing of the D1 valuation model". Medical Care. 43(3):203-20.

[11] Sullivan, P. W., Ghushchyan, V. *Mapping the EQ-5D Index from the SF-12: US General Population Preferences in a Nationally Representative Sample*. 2006. Medical Decision Making, Vol. 26, No. 4, pp. 401-409.
http://journals.sagepub.com/doi/abs/10.1177/0272989X06290496

[12] Winkler, W.E. 2006. “Overview of Record Linkage and Current Research Directions”. Research Report Series (Statistics # 2006-2). U.S. Bureau of the Census. Retrieved from http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

[13] 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