Annals of Epidemiology 24 (2014) 754–761

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

# Estimation of biomarker distributions using laboratory data collected during routine delivery of medical care

Maurice Alan Brookhart PhD [a,b,*], Jonathan V. Todd MSPH [a], Xiaojuan Li BS [a], B. Diane Reams PharmD, MPH [b], Virginia Pate MS [a], Abhijit V. Kshirsagar MD, MPH [c]

[a] Department of Epidemiology, UNC Gillings School of Global Public Health, UNC Chapel Hill, Chapel Hill, NC
[b] Cecil G. Sheps Center for Health Services Research, University of North Carolina, Chapel Hill
[c] UNC School of Medicine, University of North Carolina Kidney Center, Chapel Hill

ARTICLE INFO

*Article history:*
Received 5 January 2014
Accepted 30 July 2014
Available online 6 August 2014

*Keywords:*
National Health and Nutrition Examination Survey
Laboratories
Selection bias
Epidemiologic methods

ABSTRACT

*Purpose:* To examine the extent to which commonly ordered laboratory values obtained from large health care databases are representative of the distribution of laboratory values from the general population as reflected in the National Health and Nutrition Examination Survey.
*Methods:* Means of test values from commercial insurance laboratory data and National Health and Nutrition Examination Survey data were compared. Inverse probability of selection weighting was used to account for possible selection bias and to create comparability between the two data sources.
*Results:* The average values of most of the laboratory results from routine care were very close to their population means as estimated from NHANES. Tests that were more selectively ordered tended to differ. The inverse probability of selection weighting approach generally had a small effect on the estimated means but did improve estimation of some of the more selected tests.
*Conclusions:* Commonly ordered laboratory tests appear to be representative of values from the underlying population. This suggests that trends and other patterns in biomarker levels in the population may be reasonably studied using data collected during the routine delivery of medical care.


## Introduction

Laboratory test results are increasingly available in large health care databases and may be helpful for controlling confounding, identifying subgroups of interest, and characterizing populations. However, laboratory tests are often ordered to diagnose disease or monitor disease progress; therefore, patients with laboratory tests ordered may be a highly selected sample of the overall population. For example, patients with chronic kidney disease may be more likely to have frequent measures of serum creatinine to track renal function decline [1]. Similarly, patients with hyperlipidemia may be more likely to have lipid panels ordered so that physicians can make decisions about statin treatment [2]. Furthermore, even if a physician orders a laboratory test, it is often necessary for a patient to return later to have blood drawn. This introduces an additional selection process as many patients may fail to return for follow-up appointments. Finally, in some settings, laboratory tests are only available through certain laboratory testing companies. All factors governing the availability of laboratory values in data from routine care suggest that patients with a specific test result available may not be representative of the general population.

The selection bias created by this problem could be addressed theoretically by inverse probability of selection weighting (IPSW) [3], a semiparametric approach for missing data problems. The approach would first require a model for estimating the probability that a patient would have a laboratory test result available. The fitted model would then be used to generate IPSW that would be applied to individual patients in an analysis restricted to patients with the test result available. The approach would downweight patients who are likely to have the test result available, as they are overrepresented in the data. Similarly, patients unlikely to have a test result available would be upweighted to account for the many patients like them who are not present in the sample. The validity of the method requires that the analyst have measurements of the variables that influence both selection and the laboratory value. It is unknown how well this approach would work using typical administrative health care databases.

* Corresponding author. Department of Epidemiology, Gillings School of Global Public Health, UNC-Chapel Hill, McGavran-Greenberg, CB #7435, Chapel Hill, NC 27599-7435. Tel.: +1 (919) 843 2639; fax: +1 (919) 966 2089.
*E-mail address:* abrookhart@unc.edu (M.A. Brookhart).

1047-2797/$ – see front matter 
http://dx.doi.org/10.1016/j.annepidem.2014.07.013

Using nationally representative data from the National Health and Nutrition Examination Survey (NHANES) and routine care data from a large population of patients with commercial insurance, we examined the extent to which commonly ordered laboratory values were representative of the distribution of laboratory values from the general population as reflected in the NHANES data. We hypothesized that the values collected in routine care would be substantially biased, reflecting a population of patients with more disease and abnormal test results. We then examined the distribution of laboratory values within the IPSW sample. We hypothesized that the IPSW sample would result in distributions of laboratory results that were more comparable with those from NHANES.

## Methods

### *Data and cohort identification*

We created a cohort of patients receiving routine medical care using the Truven Health Analytics MarketScan Commercial Claims and Encounters and Laboratory databases. These databases contain individual-level information on outpatient services, procedures, diagnoses, and medication information from pharmacy records for a very large population of patients in the United States with employer-provided commercial insurance, and their dependents. This enables researchers to have longitudinal views of pharmacy and health care utilization [4]. Laboratory values are available on patients who have a test ordered and submitted to a specific national testing company. Truven links the laboratory testing data to the claims data. Using these data, we identified a retrospective cohort of patients aged 40 to 64 years who had an office visit during the calendar years 2009 to 2010 using Current Procedural Terminology codes (99211-99215, 99201-99205) to identify office visits. Next, we examined a 2-week period after the office visit to see if a patient received various laboratory tests of interest. We dropped observations in which a patient had fewer than 6 months of eligibility before the index date (date of the office visit) and less than two weeks of eligibility after the index date. To ensure that patients with multiple visits were not overrepresented in the analysis data set, we selected one physician visit at random per year to include in the analysis data set. For the selected visits, we then looked in the 2-week period after the index date to see if a patient had a specific test performed and an available result. To minimize the size of the data set, we sampled 1% of the visits with no available laboratory value.

We created a large number of covariates based on claims occurring in the 6-month period before the index date. To ensure broad capture of potentially relevant covariates, we identified all generic medications and three-digit *International Classification of Diseases, Ninth Revision* diagnostic codes that occurred with a prevalence of greater than 0.5% and procedure codes that occurred with a prevalence of greater than 1%. We created indicator variables to represent the presence of these codes. This approach is similar to other data-driven approaches to variable creation [5]. We also identified various demographic variables, such as age, sex, and region of residence.

To obtain estimates of the distribution of laboratory values from the underlying population, we used data from the 2009 to 2010 NHANES. NHANES uses probability sampling to characterize the health and nutritional status of the United States civilian population. The survey examines a sample of approximately 5000 persons each year. These persons are located in counties across the country, 15 of which are visited each year. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, and laboratory tests administered by trained medical personnel. To make the NHANES population directly comparable with our routine care population, we restricted the NHANES cohort to people aged 40 to 64 years with private insurance. People with private insurance were identified from those with a positive response to the question, "Are you/Is SP (survey participant) covered by private insurance?" as part of the household questionnaire. Appropriate examination sample weights from the mobile examination center data were used for national summary statistics of laboratory test value distributions.

### *Laboratory tests*

The laboratory tests selected for study were those commonly ordered and available in NHANES. To account for data entry errors, we worked with a clinician (A.K.) to create trimming rules for the laboratory values (Appendix Table 1). In the instance of laboratory results reported in different units, when possible, we converted those units. We examined most of the laboratories that make up the comprehensive metabolic panel including the following: electrolytes (potassium, sodium, and chloride); proteins (total protein and albumin); measures of kidney function (blood urea nitrogen and creatinine); measures of liver function (alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, and bilirubin); glucose, and calcium. We also examined components of the standard lipid panel (low-density lipoprotein, high-density lipoprotein, triglycerides, and total cholesterol), and hemoglobin concentration, white blood cell count, glycohemoglobin, C-reactive protein, and gamma-glutamyl transferase. Laboratories were identified in the routine care data using Logical Observation Identifiers Names and Codes (LOINC) codes.

### *Statistical analysis*

Using the supplied survey sampling weights, we computed summary statistics (means, standard deviations, and medians) for all selected laboratories from the NHANES cohort. These statistics were also computed for the trimmed laboratory values from routine care cohort. These statistics were then computed within strata of sex and age group. We attempted to reduce possible selection bias in the routine care cohort through IPSW. For each patient, we used a logistic model to determine the probability of having each particular laboratory value available in the 2-week period after the index visit. These models included all available covariates. Using the estimated probability of selection from the logistic regression model, we then created for each patient an inverse probability weight that was the inverse of a patient's predicted probability of having a laboratory result available given his or her observed covariate vector. Using these weights, we then computed the summary statistics of the laboratory value distribution in the reweighted sample. To diagnose possible problems with the estimated weights, we also computed the C-statistic from the fitted logistic regression model and summary statistics for the distribution of the IPSW. This process was repeated for each laboratory value.

We considered a laboratory result from routine care to be meaningfully different from the underlying population distribution if its mean was more than half a standard deviation from the NHANES mean (using the standard deviation of the laboratory result from NHANES).

All statistical analyses were performed using SAS version 9.2 software (SAS Institute, Cary, NC). The University of North Carolina Institutional Review Board approved this research.