Annals of Epidemiology 25 (2015) 526–531

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Using administrative health care system records to recruit a community-based sample for population research

J. Michael Oakes PhD *, Richard F. MacLehose PhD, Kelsey McDonald PhD, Bernard L. Harlow PhD

Division of Epidemiology, University of Minnesota, Minneapolis

ARTICLE INFO

Article history:
Received 15 September 2014
Accepted 13 March 2015
Available online 20 March 2015

Keywords:
Sampling bias
Research subject recruitment
Urban health
Methodology

ABSTRACT

Purpose: Epidemiologists often seek a representative sample of particular persons from geographically bounded areas. However, it has become increasingly difficult to identify a sample frame that truly represents the underlying target population. We assessed the degree to which a clinic-based sample represents a target community.

Methods: Our sample frame is from a large health care provider from the Minneapolis-Saint Paul, Minnesota, metropolitan area. We used U.S. Census data to examine the sociodemographic and geospatial distribution of the sampling frame and among those who did and did not respond.

Results: Our study's overall response rate was 57%. The most impoverished areas of the target population were under-represented in our sample frame, but this under-representation was similar for both respondents and nonrespondents. In addition, our sampled population was slightly older compared to the target population. Using ecological-level census-derived markers of sociodemographic characteristics, members of the sample frame were similar to that of the target population except for being somewhat more highly educated. However, the distributions of available individual-level data such as race and education were different between respondents and the target population.

Conclusions: Although the use of health care administrative records for identifying a sampling frame that represents a target population has limitations, our findings suggest that this method had strengths. More comparisons of methods for identifying and recruiting target populations are needed.

Introduction

Epidemiologists often seek a representative sample of particular persons from geographically bounded areas, such as cities or counties. Although representativeness may not be necessary for all epidemiologic investigations (see Rothman et al [1]), for some purposes, such as the surveillance in a given area, it is essential [2]. Furthermore, methods for efficiently identifying and recruiting particular subjects (e.g., by specific ages and sex) remain important for a great deal of epidemiologic research [3].

With increasing concerns over confidentiality and newly imposed limits in accessing what were once considered public records, there are fewer and fewer community-based sample frames (i.e., lists of potential participants) that are adequate for epidemiologic research [4]. Historically, commonly used options have been voter registration lists, driver's license lists, postal address lists, and various household registries. Although firm data are needed, the availability and utility of these sample frames appear to be decreasing [5,6]. For example, in 1994, the U.S. Congress adopted the Driver's Privacy Protection Act (Public Law number 103-322) that greatly restricts access to the erstwhile useful data [7], and in Minnesota, voter lists may not be used for research (MN Statute 201.091). In addition, as discussed in the following text, although random sampling of telephone numbers (random digit dialing) enjoyed early success, it is now hampered by cellular phone restrictions and the lack of geographic bounding by area code. Alternative methods for identifying and comparing community-based samples for epidemiologic research are needed.

We used the administrative records from outpatient health care facilities aligned with one of the largest health care systems in the Minneapolis-Saint Paul (MSP) metropolitan area to accrue a sample of reproductive-aged women who visited one of several outpatient clinics over the past two years without regard to reason for their visit. In this report, we compare this sample to that of the census-based data from the same target areas by responders and nonresponders and demographic characteristics. In other words, we assess the degree to which our sample sociodemographic statistics are biased with respect to population parameters.

* Corresponding author. Division of Epidemiology, 1300 South 2nd Street, Suite 300, Minneapolis, MN 55454. Tel.: +1-612-624-6855.
E-mail address: oakes007@umn.edu (J.M. Oakes).

http://dx.doi.org/10.1016/j.annepidem.2015.03.015
1047-2797/© 2015 Elsevier Inc. All rights reserved.

Methods

As part of a larger study on the prevalence and etiology of unexplained vulvar pain (vulvodynia), we aimed to recruit a sociodemographically representative sample of women aged 18–40 years from the MSP metropolitan area of Minnesota. Because there is no publicly available sample frame of all women age 18–40 years in this geographic region, we decided to rely on administrative records from an affiliated health care system. The affiliated organization is a not-for-profit fee-for-service system that encompasses many hospitals, an academic health center, and many outpatient clinics, including clinics known to primarily provide care to disadvantaged and uninsured populations. As with many health care providers today, our affiliated system is a large organization that provides health care services to approximately 25% of the area's 2.5 million people and uses a uniform electronic medical record system.

With Institutional Review Board approval, we obtained data on the name, age, and home address of women within this age range who visited one of five purposively selected local outpatient clinics at least once for any reason within the previous 24 months of each of two administrative database acquisitions (outpatient clinic visits between March, 2008 and July, 2011). The five clinics are large, well-known primary care clinics in targeted geographic area and serve a diverse population, both economically and ethnically. Because we wanted a subset of selected women to accept our invitation to return to their (or a nearby affiliated) clinic for a gynecological examination, we limited our focus to women who actually visited one of the five local clinics, thus excluding persons who only visited an affiliated emergency department or other specialty (e.g., orthopedic) clinic in the health care system. No diagnostic code filters were applied and no other medical record information was attached; we simply obtained this sample frame for recruiting a sample of women aged 18–40 years. The name and address information from the most recent visit was used. As usual, no socioeconomic (e.g., educational attainment) measure was recorded or included in the administrative data, and the race and ethnicity variable was largely missing and deemed unreliable (refer the Discussion section). We used a geographic information system to geocode patient addresses and excluded those residing more than 65 miles from the centroid of the MSP area. From this administrative database sampling frame, we drew a simple random sample of 14,321 and sent these women a study invitation letter, $2.00 cash, an "opt-out" postcard, and a short double-sided one page health screening survey via U.S. mail. Our procedures were modeled after Dillman's Tailored Design Method [8]. Nonrespondents were mailed a second and third letter and screener at approximately 4 and then 8 weeks after the initial mailing, respectively.

We undertook two separate analyses to estimate how similar the sample of respondents in our study was to the target population. First, we conducted an ecological analysis using the geocoded data from each woman in the sample (participants and nonparticipants) to her 2010 census block group and assigned her the block group's sociodemographic characteristics for race, education, income, and poverty, which were compiled by and downloaded from the Minnesota Population Center's National Health Geographic Information System [9]. Although this is imperfect and subject to an ecological fallacy [10], given that there is well-known residential racial and socioeconomic segregation in the MSP target population, there is reason to support this approach [11,12]. The neighborhood in which people live is often a reflection of who they are sociodemographically [13]. A second analysis was conducted comparing responders to the target population on those variables for which individual-level information was available (race and education). Census data were used to determine the distribution of race and education in the target population (as in the first analysis). The screener questionnaire was used to determine the distribution of these variables among respondents.

Table 1
Ecological-level demographic comparisons

Demographic	Target	Sample	Participants	Non participants
White persons (%)	64.58	63.57	64.48[*]	62.37
Bachelor's degree or higher (%)	34.67	42.84	44.28[†]	40.92
Median household income, $1000	66.19	68.44	68.56	68.27
Households below poverty line (%)	10.71	9.73	9.53	9.99

[*] Statistic calculated from individual level data is 85.7%.
[†] Statistic calculated from individual level data is 59.2%.

We used data from the 2010 U.S. Decennial Census and the Census Bureau's American Community Survey pooled from 2006 to 2010, again from the National Health Geographic Information System, to compare our sample frame and respondents to the target area's demographic characteristics. Although not without flaws, these reference data are considered the gold standard. For purposes here, only simple tabulations and descriptive statistics are calculated. The Decennial Census (SF1) was used to compile age group and race variables; however, as they were not available in the decennial data, the American Community Survey was used to compile all other census variables in Table 1. Although we acknowledge that the American Community Survey is a complex sample survey, given the sample size of the census data used, we treat such data as if sampling variability is negligible. Accordingly, statistical tests of significance are not appropriate here. All analyses were conducted with Stata SE, version 12.1 (Stata Statistical Software, College Station, TX), and ArcGIS, version 10.1 (ArcGIS Desktop, Redlands, CA).

Results

We first assessed the sociodemographic representativeness of the respondents selected from the health care administrative database to determine whether our health care administrative data sample is representative of the target population (based on census data). Table 2 compares the age distributions of the geographical target area (from census data), our sample of all subjects surveyed, respondents, and nonrespondents. After exclusion of those deemed ineligible because of incorrect addresses or age restrictions within our random sample of 14,321 women, the overall participation rate for this study was 57%. According to the census, there were 571,130 females aged 18–39 years in the target area in 2009 (for the 2010 census tabulations). The age distribution from our sample is derived from the medical records themselves (the Pearson correlation between self-reported and medical record age is $r = 0.99$). Our sample mean age (30.59 years) was slightly greater than the census mean

Table 2
Age distribution of eligible women in target area

Age, y	Census data		Medical record data					
	Study area		Sample frame		Participants		Non participants	
	N	%	*N*	%	*N*	%	*N*	%
Total	571,130		14,321		8179		6142	
18–19	50,266	8.80	346	2.42	138	1.69	208	3.39
20–21	49,947	8.75	818	5.71	333	4.07	485	7.90
22–24	78,275	13.71	1416	9.89	711	8.69	705	11.48
25–29	141,374	24.75	3891	27.17	2266	27.71	1625	26.46
30–34	128,236	22.45	4163	29.07	2544	31.10	1619	26.36
35–39	123,032	21.54	3687	25.75	2187	26.74	1500	24.42