Published in final edited form as:

Ann Epidemiol. 2020 October ; 50: 48–51.e2. doi:10.1016/j.annepidem.2020.06.008.

Reweighting to address non-participation and missing data bias in a longitudinal electronic health record study

Milena A. Gianfrancesco,

Division of Rheumatology, Department of Medicine, University of California, San Francisco;

Charles E. McCulloch,

Department of Epidemiology and Biostatistics, University of California, San Francisco;

Laura Trupin,

Division of Rheumatology, Department of Medicine, University of California, San Francisco;

Jonathan Graf,

Division of Rheumatology, Department of Medicine, University of California, San Francisco;

Gabriela Schmajuk,

Division of Rheumatology, Department of Medicine, University of California, San Francisco;
Veterans Affairs Medical Center, San Francisco;

Jinoos Yazdany

Division of Rheumatology, Department of Medicine, University of California, San Francisco;

Abstract

Purpose: We examined whether weighting techniques could account for longitudinal differences in disease activity by race/ethnicity between research participants and non-participants with rheumatoid arthritis (RA).

Methods: We included 377 RA patients from a public hospital in San Francisco, CA. We estimated the probability of not enrolling in a research study by constructing weights using inverse probability weighting (IPW). Disease activity over time by race/ethnicity was analyzed across the

*Corresponding author and requests for reprints: Milena Gianfrancesco PhD MPH; Division of Rheumatology, Department of Medicine, University of California, San Francisco; 513 Parnassus Avenue, San Francisco, CA 94143; Milena.Gianfrancesco@ucsf.edu.

Milena A. Gianfrancesco: Writing-Original draft presentation; Conceptualization; Methodology; Formal Analysis;

Charles E. McCulloch: Writing-Reviewing and Editing; Methodology

Laura Trupin: Writing-Reviewing and Editing; Data curation

Jonathan Graf: Writing-Reviewing and Editing; Resources

Gabriela Schmajuk: Writing-Reviewing and Editing; Supervision

Jinoos Yazdany: Writing-Reviewing and Editing; Supervision; Resources

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of Interest:

There are no conflicts of interest with respect to this study. Drs. Yazdany and Schmajuk have received an independent research award from Pfizer.

entire patient population, and amongst research participants only using multivariable mixed effects models.

Results: There were no differences in RA disease activity scores between research participants and non-participants at baseline; however, longitudinal differences in disease activity between research participants and non-participants were found by race/ethnicity. Weighting research participants according to sociodemographic and clinical characteristics of the non-participant population did not result in any meaningful changes in disease activity by race/ethnicity over time.

Conclusions: In our study of RA patients, IPW using select sociodemographic and clinical variables was not sufficient to account for longitudinal disease activity differences by race/ethnicity between research participants and non-participants.

Keywords

selection bias; electronic health records; inverse probability weighting; generalizability

BACKGROUND

A well-designed epidemiologic study rests on the assumption that the study sample is representative of the target population. However, decreasing rates of participation in research over the last few decades have been described,¹ potentially threatening the generalizability and validity of study findings. Further, participation rates have been shown to vary by factors such as sex,² age,³ and race/ethnicity,^{4,5} which may introduce selection bias and lead to inaccurate understanding of disease processes for certain subgroups.

While several studies have attempted to illustrate the representativeness of a given research population by comparing enrolled subjects to those who decline to participate, or in comparison to the corresponding census population, these approaches miss the opportunity to understand long-term differences in disease outcomes between groups over time. For example, even when differences in demographic and clinical factors may not be significant between two populations at baseline, there still may be marked differences in longitudinal outcomes.⁶⁻⁹

In a previous study, we found substantial differences in longitudinal disease activity trajectories between research participants and non-participants in a population of patients with rheumatoid arthritis (RA).⁹ Specifically, racial/ethnic disparities in RA disease activity over time were present in non-participants, but not in research participants, potentially limiting our understanding of how disease may be influenced by race/ethnicity if studying research participants only.

Statistical strategies, such as weighting, have been applied to research studies to represent the population of interest as closely as possible.^{10,11} One such technique, inverse probability weighting (IPW), can be used to weight the probability of participation in order to control for potential responder bias based on certain variables. This technique is appropriate for longitudinal designs.^{10,12}

Taking advantage of a unique study cohort in which all patients attending a clinic were invited to participate, this study sought to examine whether weighting according to the probability of participating in a research study by sociodemographic and clinical factors using IPW could explain longitudinal differences in disease activity by race/ethnicity between participants and non-participants with RA, and more closely approximate estimates found in the combined patient population.

METHODS

Data source and study population

We included individuals with a diagnosis of RA (ICD-9 diagnosis code: 714.0) and at least two rheumatology clinic visits within 12 months between January 1, 2013 – February 28, 2017 from the EHR of a public hospital in San Francisco, CA. Our study population included 377 patients with 2,269 documented EHR visits.

Details of the study design, including enrollment and data collection, have been previously reported.^{9,13–16} Briefly, all patients with RA seen in the health system's rheumatology clinic were approached to participate in the Rheumatoid Arthritis Observational Cohort Study during their initial or return visits. Patients who agreed to participate comprised the "participant" subset in our analyses (n=187 individuals, 1,491 observations). Data from research participants were collected during regularly scheduled clinic visits and documented in the EHR. Additional information collected through questionnaires was available in a separate database. Patients who did not agree to participate comprised the "non-participants," and included 190 patients (n=778 observations) with data extracted exclusively from the EHR. All patients were observed from first eligible visit (second of two encounters within 12 months) until loss to follow-up or February 28, 2017, whichever occurred first.

Data collection

Data utilized for this study included information extracted from the EHR, including sex, race/ethnicity (white non-Hispanic; black non-Hispanic; Asian non-Hispanic; Latino/Hispanic or other race), language, age, body mass index, smoking status (ever/never), whether a patient was prescribed a disease modifying anti-rheumatic drug (DMARD), coded as conventional synthetic (csDMARD) (yes/no) and biologic or targeted small molecule (bDMARD) (yes/no), and total number of rheumatology clinic visits. In an attempt to define a marker of external social circumstances preventing engagement in care,^{17–19} we calculated the proportion of "no-show" visits scheduled at the public hospital across all clinics for each patient. This was done by summing the number of in-person visits a patient failed to attend without cancelling in advance and dividing it by the total number of in-person visits over the study period. If two visits occurred on the same day in different clinics, they were counted separately. Our primary outcome was disease activity measured by the Clinical Disease Activity Index (CDAI),²⁰ a composite measure of RA disease activity that includes a tender and swollen joint count, as well as a patient-reported outcome (patient global), and physician assessment (physician global). We also conducted analyses examining the patient and physician global subscores as independent outcomes.

All covariates contained less than 6% missing values. Not all participants had disease activity scores recorded at every visit, resulting in a final sample size of 1,297 observations for CDAI, 2,088 for patient global, and 1,768 for physician global scores.

Statistical analyses

In order to account for differences between research participants and non-participants, we estimated the probability of *not* enrolling in the research study (e.g. the “non-participant” population), and constructed weights using IPW for all individuals.¹² We created two different weightings: first, we calculated the probability of not enrolling in the research study using logistic regression, accounting for sociodemographic and clinical variables [$P(A/X)$], where A represents whether a patient was a non-participant ($A=1$) or participant ($A=0$), and X represents variables: age, sex, race/ethnicity, baseline disease activity, smoking status, body mass index, language, csDMARD, bDMARD, and total number of rheumatology clinic visits. (“weighting 1”). Second, we calculated the probability of not enrolling in the research study accounting for sociodemographic and clinical variables as well as the proportion of “no-show” visits during the study period (“weighting 2”). We then calculated weights for all participants according to the following equation: $\frac{I_A}{p(A|X)} + \frac{1 - I_A}{1 - p(A|X)}$, where I is 1 for participants and 0 for non-participants. This weighting “inflates” each of the participant and non-participant subgroups to the combined cohort.

Differences in longitudinal disease activity by race/ethnicity over the study period were analyzed across the entire EHR population (research participants and non-participants), and amongst research participants only using multivariable mixed effects models, controlling for baseline (sex, age, body mass index, smoking status) and time-varying (csDMARD medication, bDMARD medication, and time as measured by days since last visit) covariates. Weights were included in the models when specified. Cluster robust standard errors were used to accommodate the sampling weights and avoid making assumptions about the variance-covariance structure. The distribution of weights by race/ethnicity using kernel density plots illustrated satisfactory overlap. A sensitivity analysis using stabilized weights based on the 1st and 99th percentile, as well as the 5th and 95th percentile, were also conducted.

We created 95% confidence intervals (CI) and conducted two-sided hypothesis tests controlling the type I error rate at 5% ($\alpha = 0.05$). Analyses were conducted in Stata (v.15.0). The study was approved by the Committee on Human Research at the University of California, San Francisco.

RESULTS

Sociodemographic and disease characteristics of patients by participation are shown in Table 1, and by race/ethnicity in eTable 1. There were no differences in RA disease activity as measured by CDAI between research participants and non-participants at baseline. However, there was a difference in follow-up time and no-show visits between these groups over the study period; on average, research participants were labeled a “no-show” at 14% of visits, while non-participants were labeled a “no-show” at 18% of visits during the study period.

Differences in disease activity by racial/ethnic groups amongst the entire EHR population were substantially different from those found within research participants only (Table 2). After weighting the research participant population according to the non-participant population in terms of sociodemographic and clinical characteristics (weighting 1), we did not find any meaningful changes in disease activity by race/ethnicity. For example, black, non-Hispanic race/ethnicity was associated with a beta of -0.40 and -1.91 in the unweighted and weighted participant population, respectively, versus 3.77 in the analysis including both research participants and non-participants.

Similar results were seen when examining the patient and physician global subscores of disease activity (eTables 2 and 3). Reweighting the research participants according to sociodemographic and disease characteristics of the non-participants did not result in values reflecting differences in disease activity score by race/ethnicity seen in the combined analysis of both research participants and non-participants.

Additional weighting accounting for the proportion of no-show visits (weighting 2) across all three analyses had no meaningful effect. Both models demonstrated acceptable performance (AUC = 0.76 for weighted model 1, and 0.77 for weighted model 2; eTable 4). Results were identical using 1st and 99th percentile stabilized weights, and remained essentially unchanged using 5th and 95th percentile stabilized weights (data not shown).

CONCLUSIONS

In this study, we show that using IPW based on EHR derived sociodemographic and clinical characteristics to account for differences in disease activity by race/ethnicity between research participants and non-participants did not yield results that match analyses of the entire, combined EHR population. In our example, we significantly underestimate the effect of race/ethnicity on longitudinal disease activity for minority groups if we only consider the participant population, even after reweighting using available measures.

In order to truly understand disease at a population level, it is essential to include a study sample representative of the target population, though realistically this is not always possible. Weighting methods, such as IPW, have been used to account for non-response bias,²¹ but rely on correct specification of the selection model. We hypothesize that there are likely additional factors that are not captured in the EHR that explain differences in disease activity between research participants and non-participants, such as adherence to medications and social determinants of health, that were not represented in the IPW model. While we attempted to include a proxy for likely external social circumstances that may prevent engagement in care (proportion of “no-show” visits), which differed between research participants and non-participants, this variable did not result in any meaningful changes in the longitudinal analyses. Results from this study are limited by the small sample size and may not be generalizable to other conditions or RA patients in different health care settings. We also were unable to access clinical notes, which may have contained more information for uncontrolled confounders, such as social determinants of health.

In summary, we found that IPW using available sociodemographic and clinical variables was not sufficient to account for longitudinal disease activity differences by race/ethnicity between research participants and non-participants with RA. Given these results, researchers must be aware of the potential limitations and lack of generalizability of findings in study designs that do not address participation bias, or fail to correctly specify all contributing factors to the model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

This work was supported by grants K01 AR070585 (MAG) and K24 AR074534 (JY) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health. Drs. Yazdany and Schmajuk are also supported by the Russell/Engleman Medical Research Center for Arthritis. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Galea S, Tracy M. Participation rates in epidemiologic studies. *Ann Epidemiol.* 2007;17(9):643–653. [PubMed: 17553702]
2. Markanday S, Brennan SL, Gould H, Pasco JA. Sex-differences in reasons for non-participation at recruitment: Geelong Osteoporosis Study. *BMC Res Notes.* 2013;6:104. [PubMed: 23506528]
3. Golomb BA, Chan VT, Evans MA, Koperski S, White HL, Criqui MH. The older the better: are elderly study participants more non-representative? A cross-sectional analysis of clinical trial and observational study samples. *BMJ Open.* 2012;2(6).
4. Falasinnu T, Chaichian Y, Bass MB, Simard JF. The Representation of Gender and Race/Ethnic Groups in Randomized Clinical Trials of Individuals with Systemic Lupus Erythematosus. *Curr Rheumatol Rep.* 2018;20(4):20. [PubMed: 29550947]
5. Garza MA, Quinn SC, Li Y, et al. The Influence of Race and Ethnicity on Becoming a Human Subject: Factors Associated with Participation in Research. *Contemp Clin Trials Commun.* 2017;7:57–63. [PubMed: 29226266]
6. Yeh RW, Kennedy K, Spertus JA, et al. Do postmarketing surveillance studies represent real-world populations? A comparison of patient characteristics and outcomes after carotid artery stenting. *Circulation.* 2011;123(13):1384–1390. [PubMed: 21422383]
7. Manjer J, Carlsson S, Elmstahl S, et al. The Malmo Diet and Cancer Study: representativity, cancer incidence and mortality in participants and non-participants. *Eur J Cancer Prev.* 2001;10(6):489–499. [PubMed: 11916347]
8. Drivsholm T, Eplov LF, Davidsen M, et al. Representativeness in population-based studies: a detailed description of non-response in a Danish cohort study. *Scand J Public Health.* 2006;34(6):623–631. [PubMed: 17132596]
9. Gianfrancesco MA, Trupin L, McCulloch CE, Shiboski S, Schmajuk G, Yazdany J. Differences in Longitudinal Disease Activity Between Research Cohort and Noncohort Participants with Rheumatoid Arthritis Using Electronic Health Record Data. *ACR Open Rheumatol.* 2019;1(2):113–118. [PubMed: 31777787]
10. Schmidt SCE, Woll A. Longitudinal drop-out and weighting against its bias. *BMC Med Res Methodol.* 2017;17(1):164. [PubMed: 29221434]
11. Kalton G, Flores-Cervantes I. Weighting Methods. *J Off Stat.* 2003;19(2):81–97.
12. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. *J Am Stat Assoc.* 1952;47(260):663–685.

13. Barton JL, Imboden J, Graf J, Glidden D, Yelin EH, Schillinger D. Patient-physician discordance in assessments of global disease severity in rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62(6):857–864. [PubMed: 20535797]
14. Barton JL, Trupin L, Schillinger D, et al. Racial and ethnic disparities in disease activity and function among persons with rheumatoid arthritis from university-affiliated clinics. *Arthritis Care Res (Hoboken)*. 2011;63(9):1238–1246. [PubMed: 21671414]
15. Graf J, Scherzer R, Grunfeld C, Imboden J. Levels of C-reactive protein associated with high and very high cardiovascular risk are prevalent in patients with rheumatoid arthritis. *PLoS One*. 2009;4(7):e6242. [PubMed: 19606218]
16. Margaretten M, Barton J, Julian L, et al. Socioeconomic determinants of disability and depression in patients with rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2011;63(2):240–246. [PubMed: 20824800]
17. Coodin S, Staley D, Cortens B, Desrochers R, McLandress S. Patient factors associated with missed appointments in persons with schizophrenia. *Can J Psychiatry*. 2004;49(2):145–148. [PubMed: 15065750]
18. Mason C Non-attendance at out-patient clinics: a case study. *J Adv Nurs*. 1992;17(5):554–560. [PubMed: 1602070]
19. Campbell B, Staley D, Matas M. Who misses appointments? An empirical analysis. *Can J Psychiatry*. 1991;36(3):223–225. [PubMed: 2059940]
20. Aletaha D, Nell VP, Stamm T, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. *Arthritis Res Ther*. 2005;7(4):R796–806. [PubMed: 15987481]
21. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. *Acta Obstet Gynecol Scand*. 2018;97(4):407–416. [PubMed: 29415329]

Table 1.
Baseline demographic and disease characteristics of participants and non-participants with rheumatoid arthritis in an electronic health record database

	Participants (n=187)	Non-Participants (n=190)
Female sex	158 (84%)	143 (75%)
Age	59.7 (12.6)	57.0 (12.6)
Race/Ethnicity		
White, non-Hispanic	12 (6%)	21 (11%)
Asian	57 (30.5%)	51 (27%)
Black, non-Hispanic	12 (6%)	21 (11%)
Latino/Hispanic or other race	106 (57%)	97 (51%)
Language		
English	64 (34%)	76 (40%)
Spanish	69 (37%)	53 (28%)
Cantonese	38 (20%)	18 (9%)
Other	16 (9%)	16 (8%)
Unknown	0 (0%)	27 (14%)
Body mass index	28.4 (6.2)	29.8 (7.9)
Ever smoker	13 (7%)	20 (11%)
Conventional synthetic DMARD	138 (74%)	120 (64%)
Biologic/small molecule DMARD	70 (38%)	20 (11%)
Number of visits/person	8.0 (3.5)	4.1 (3.1)
Follow-up time (years)	2.10 (0.91)	0.99 (0.96)
Clinical Disease Activity Score (range 0–76)	14.6 (11.8)	15.1 (12.9)

DMARD = disease-modifying anti-rheumatic drug

Longitudinal differences in Clinical Disease Activity Index (CDAI) score (0–76) by race/ethnicity using unweighted and weighted mixed-effects models

Table 2.

Race/Ethnicity	All Participants (Unweighted) N=1943 visits, N=339 participants		Cohort Participants (Unweighted) N=1297 visits, N=177 participants		Cohort Participant (Weighted 1*) N=1297 visits, N=177 participants		Cohort Participants (Weighted 2*) N=1297 visits, N=177 participants	
	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
White, non-Hispanic	(Ref)		(Ref)		(Ref)		(Ref)	
Asian	-0.87	1.54	-2.96	2.03	-4.95	2.86	-5.05	2.84
Black, non-Hispanic	3.77	1.89	-0.40	2.61	-1.91	3.06	-1.99	3.05
Latino/Hispanic or Other Race	3.35	1.42	1.89	1.85	0.28	2.58	0.17	2.55

Mixed effects model over 38-month period; controlled for age, sex, body mass index, smoking status, conventional synthetic DMARD, biologic/small molecule DMARD, and time (days since last visit).

Weighted 1: research participants weighted to non-participants according to age, sex, race/ethnicity, baseline disease activity, smoking status, body mass index, language, conventional synthetic DMARD, biologic/small molecule DMARD, and total number of visits.

Weighted 2: research participants weighted to non-cohort according to "Weighted 1" covariates and proportion of no-show visits.