Differences in Longitudinal Disease Activity Between Research Cohort and Noncohort Participants with Rheumatoid Arthritis Using Electronic Health Record Data

Milena A. Gianfrancesco, Laura Trupin, Charles E. McCulloch, Stephen Shiboski, Gabriela Schmajuk and Jinoos Yazdany

Objective. Research using electronic health records (EHRs) may offer advantages over observational prospective cohort studies, including lower costs and a more generalizable patient population; however, EHR data may be more biased because of the high prevalence of missing data. We took advantage of a unique clinical setting in which all patients with rheumatoid arthritis (RA) were asked to participate in a longitudinal cohort study that would examine potential biases of EHR vs. prospective cohort designs in assessment of disease outcomes, but only some chose to participate.

Methods. For individuals both participating in the cohort (“cohort,” n = 187) and not participating (“noncohort,” n = 190), we retrieved data regarding RA disease activity and other sociodemographic and clinical factors from data recorded in the EHR between 2013 and 2017. We compared the prevalence of missing data between groups and studied differences in disease activity measures over time.

Results. Disease activity measures were less likely to be missing for cohort participants compared with noncohort participants (0.2%-13% vs. 2%-22%, respectively). No significant differences were present at baseline with respect to race/ethnicity or disease activity measures between groups. However, black, non-Hispanic race/ethnicity was associated with worse longitudinal disease activity compared with white, non-Hispanic individuals in noncohort participants (β = 6.47, P = 0.03) but not in cohort participants (β = −0.10, P = 0.97) (P interaction = 0.09).

Conclusion. Findings suggest that data derived from the EHR were comparable to a cohort across some variables but captured racial/ethnic disparities in long-term outcomes not observed in the cohort study. Research utilizing EHR data in conjunction with cohort studies may provide new opportunities for studying health disparities.

INTRODUCTION

Advances in electronic health records (EHRs) have allowed researchers to access clinical data generated during routine care that permits exploration of hypotheses with lower overhead costs and without some of the burdens of study recruitment (1). Although cohort studies have the advantage of comprehensively recording a number of different factors using validated research protocols, these studies are often met with small sample sizes and/or high costs. EHRs offer the advantage of simultaneously examining multiple risk factors and outcomes with great power in a relatively short amount of time. A recent study found that well-established epidemiologic features of multiple sclerosis natural history could be extracted and replicated using EHR clinical data (2). It has also been suggested that EHR analyses allow for greater sample sizes of a representative patient population (1), albeit one that seeks medical care.
Although informal comparisons between cohort participants and EHR-derived populations have been described in the literature, formal comparisons of longitudinal outcomes between cohort and noncohort participants from the same underlying population have not been extensively explored. Rates of refusal and demographic characteristics of those who do and do not participate in studies give researchers an idea of potential differences between participants and nonparticipants; however, long-term disease trajectories of nonparticipants are generally unknown. Research studies may be particularly susceptible to selection bias, specifically volunteer bias, which occurs when individuals who volunteer to participate in research are different from the general population. For example, patients from vulnerable populations may not elect to participate in longitudinal research studies. Difficulty in recruitment of minority groups has been cited as an issue in randomized controlled trials (3) and cohort studies (4), including those studying rheumatic diseases. A recent review found that while African Americans represent approximately 40% of systemic lupus erythematosus cases, they accounted for only 14% of randomized controlled trial enrollees (5). If patients from vulnerable populations or those with severe disease disproportionately do not participate in various types of studies, an important segment of the population will be missing from research that aims to predict disease trajectories of important patient outcomes.

The goal of this study was to examine differences in baseline demographics and disease outcomes between cohort and noncohort participants from a population of rheumatoid arthritis (RA) patients captured in the EHR of a single health system. We also compared the availability of measures and prevalence of missing data between the two groups and explored differences in longitudinal predictors of RA disease activity.

PATIENTS AND METHODS

Study population. We included individuals with a diagnosis of RA (International Classification of Disease-9 diagnosis code: 714.0) and at least two face-to-face rheumatology clinic visits within 12 months between January 1, 2013, and February 28, 2017, from the EHR of a public hospital in San Francisco, California. This resulted in a study population of 377 unique patients with a total of 2269 documented EHR visits.

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 (6). These participants comprise the cohort subset in our analyses. Enrollment and data collection details have been previously reported; briefly, eligible patients were 18 years or older and met the American College of Rheumatology classification criteria for RA (7). Data from cohort participants were collected during regularly scheduled clinic visits and documented in the EHR. Additional information collected as part of cohort visits through questionnaires, including patient-reported outcomes, was available in a separate database. A total of 187 individuals were included as cohort participants and accounted for approximately 66% of all EHR visits (n = 1491 observations). The remainder of the 377 patients are what we describe as the noncohort subset (n = 190 patients, 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. We extracted variables of interest, including sociodemographic, clinical, and RA outcome measures to compare availability of different measures captured in the cohort study vs. EHR (Supplementary Table 1). Covariates for longitudinal analyses included sex, race/ethnicity (white non-Hispanic; black non-Hispanic; Asian non-Hispanic; Hispanic), age, body mass index, smoking status (ever/never), and disease-modifying anti-rheumatic medications (DMARDs) coded as synthetic DMARD (yes/no) and biologic/small molecule DMARD (yes/no). Outcomes included disease activity measured by the Clinical Disease Activity Index (CDAI) (8), a composite measure of RA disease activity that includes a tender and swollen joint count as well as patient and physician global assessments. For both groups, the CDAI score was derived from a structured field as part of the rheumatology EHR template. Starting in 2013, educational efforts with providers and trainees began with the goal of increasing collection of disease activity scores across clinical sites at the University of California, San Francisco. These initiatives in this study were primarily focused on collection at the university clinic rather than the public clinic (and included templates, online homunculus tools to help with joint counts, and regular performance reports), but many of the providers at the university and the county clinic are the same. A structured field was added to the EHR in the county clinic to improve CDAI capture in early 2016. In addition, beginning in 2016, the county clinic participated in an incentive program that provided additional funds to the clinic if certain targets, set by the providers themselves, were met. Completion of CDAI at every visit was one of the targeted measures. Together, these educational, workflow, and incentive modifications resulted in high rates of CDAI capture.

Statistical analyses. We compared baseline characteristics of cohort and noncohort participants using \( \chi^2 \) tests for categorical variables and \( t \) tests for numeric variables. We also calculated missing data percentages for demographic and disease characteristics, as well as various outcome measures for each group.

In order to examine whether differences in longitudinal disease activity were present between cohort and noncohort groups, multivariate mixed effects models were used to examine the association between covariates and the CDAI score over the study period. In a sensitivity analysis, the CDAI score was also modeled with a square root-transformation because of its nonnormal distribution of scores and residuals. We utilized a complete-case analysis for our models, adjusting for baseline covariates—sex, race/ethnicity, age,
body mass index, smoking status—and time-varying covariates—synthetic DMARD medication, biologic/small molecule DMARD medication, and days since last visit. Interaction terms were additionally created and included in the multivariate model to test for interaction between each covariate predictor and participant status (cohort participant vs. noncohort participant). As a sensitivity analysis, we imputed missing data with the multivariate imputation using chained equations (MICE) method. We also conducted analyses with the additional criteria that all patients be prescribed at least one DMARD at any time point during the study period so as to decrease the possibility that non-RA patients were included in our analysis.

Analyses were conducted in Stata (v.15.0). We created 95% confidence intervals (CIs) and conducted two-sided hypothesis tests controlling the type I error rate at 5% (α = 0.05). Because we were interested in detecting suggestive interactive effects between covariates and participant status, interaction P values were considered significant if they were less than 0.10. The study was approved by the Committee on Human Research at the University of California, San Francisco.

### RESULTS

Baseline demographic and disease characteristics of RA patients included in analyses are outlined in Table 1. Cohort participants (n = 187) were significantly different from noncohort (n = 190) participants with respect to sex, age, and preferred language (P < 0.05). Notably, 14% of noncohort participants did not have information on language (ie, "unknown") compared with 0% of cohort participants. There was a higher percentage of Spanish and Cantonese speakers in the cohort group, reflecting the availability of questionnaires and interviewers for these languages. Noncohort participants were less likely to be prescribed a biologic/small molecule or synthetic DMARD at baseline, and had fewer overall visits than cohort participants (P < 0.05). No differences in overall CDAI disease activity scores or subcomponents were found.

Supplementary Table 1 compares the percentage of missing data present in both EHR and RA cohort databases across various variables for both cohort and noncohort groups. Fixed

### Table 1. Demographic and baseline disease characteristics of RA cohort and noncohort participants from the EHR of a public hospital in San Francisco, CA, from 2013-2017

|                      | Cohort Participants (n = 187)\(^a\) | Noncohort Participants (n = 190)\(^a\) | \(P\) value |
|----------------------|------------------------------------|---------------------------------------|------------|
| Sex (female)         | 158 (75%)                          | 143 (84%)                             | 0.03       |
| Age                  | 59.65 (12.63)                      | 56.96 (12.57)                         | 0.04       |
| Race/Ethnicity       |                                    |                                       |           |
| White, non-Hispanic  | 12 (6%)                            | 21 (11%)                              | 0.21       |
| Asian/Pacific Islander | 57 (30%)                          | 51 (27%)                              |           |
| Black, non-Hispanic  | 12 (6%)                            | 21 (11%)                              |           |
| Hispanic             | 104 (56%)                          | 94 (49%)                              |           |
| Other/mixed race     | 2 (1%)                             | 3 (2%)                                |           |
| Language             |                                    |                                       | <0.001     |
| English              | 64 (34%)                           | 76 (40%)                              |           |
| Spanish              | 69 (37%)                           | 53 (28%)                              |           |
| Chinese - Cantonese  | 38 (20%)                           | 18 (10%)                              |           |
| Other                | 16 (9%)                            | 16 (8%)                               |           |
| Unknown              | 0 (0%)                             | 27 (14%)                              |           |
| Body mass index      | 28.38 (6.19)                       | 29.81 (7.90)                          | 0.07       |
| Current smoker       | 20 (11%)                           | 13 (7%)                               | 0.21       |
| Biologic/small molecule DMARD | 70 (38%)                     | 20 (11%)                              | <0.001     |
| Synthetic DMARD      | 138 (74%)                          | 120 (64%)                             | 0.04       |
| Clinical Disease Activity Index (0-76) | 14.86 (11.92)                     | 15.36 (13.14)                        | 0.71       |
| Patient Global score (0-10) | 4.72 (2.57)                     | 5.16 (3.00)                           | 0.13       |
| Physician Global score (0-10) | 2.69 (2.31)                      | 2.60 (2.49)                           | 0.76       |
| Swollen joint count (0-28) | 4.25 (5.23)                     | 4.26 (5.97)                           | 1.00       |
| Tender joint count (0-28) | 3.16 (5.03)                     | 4.17 (6.13)                           | 0.10       |
| Number of visits/person | 7.97 (3.49)                     | 4.09 (3.15)                           | <0.001     |

\(^a\) Table values represent: N (%) or mean (SD).
variables, such as sex, age, and race/ethnicity, were completely available for both groups. However, language was more likely to be missing in noncohort participants compared with cohort participants, likely due to it being a requirement for inclusion. In addition, there were differences in the percentage of missing variables collected over time. For instance, Physician Global score (13% vs. 22%), swollen joint count (11% vs. 21%), and tender joint count (11% vs. 21%) were less likely to be missing for cohort participants compared with noncohort participants. Labs such as erythrocyte sedimentation rate and C-reactive protein were slightly more likely to be missing for cohort participants (23% vs. 17% and 18% vs. 13%, respectively). Furthermore, certain variables were not available for noncohort participants because they are not routinely collected and/or documented in the EHR, such as biological samples (ie, for genetic analyses) time period. To our knowledge, this is the first study to compare differences in variable availability and longitudinal predictors of RA disease activity between cohort and noncohort participants from the same underlying EHR population. Unlike disparities shown in recruitment of participants into randomized controlled trials (5), we found no significant differences between cohort and noncohort participants at baseline with respect to race/ethnicity or disease activity measures. This suggests that there was no meaningful selection bias of participants based on individual factors such as sex, age, and race/ethnicity not taking a synthetic DMARD ($\beta = 1.25$, $P = 0.27$), whereas the inverse was shown in the cohort participant group ($\beta = -1.38$, $P = 0.05$) ($P$ interaction = 0.07). Imputation analyses, as well as analyses using transformed values for CDAI scores, demonstrated similar results (data not shown).

Analyses requiring the additional criteria of being on at least one DMARD at any time point during the study period showed consistent findings. In noncohort participants, black, non-Hispanic race/ethnicity was associated with a significantly higher CDAI score compared with white, non-Hispanic individuals ($\beta = 8.30$, $P = 0.02$), but this was not found for cohort participants ($\beta = 1.76$, $P = 0.52$) ($P$ interaction = 0.13). Differences by synthetic DMARD status were not able to be assessed because they were conditioned on being prescribed at least one DMARD during the time period.

**DISCUSSION**

This study illustrates how the addition of EHR data, which may include noncohort individuals, to cohort studies may increase generalizability of the patient population and understanding of disease trajectories. To our knowledge, this is the first study to compare differences in variable availability and longitudinal predictors of RA disease activity between cohort and noncohort participants from the same underlying EHR population.

Table 2. Multivariate analysis of predictive factors on Clinical Disease Activity Index (CDAI) scores in RA cohort and noncohort participants

| Variable                        | Cohort Participants (N = 1,337 visits) | Noncohort Participants (N = 625 visits) |
|---------------------------------|----------------------------------------|----------------------------------------|
|                                 | $\beta$  | 95% CI        | $P$ value | $\beta$  | 95% CI        | $P$ value | $P$ Interaction |
| Female sex                      | 0.64     | -2.03, 3.32   | 0.64      | 2.84     | -0.85, 6.54   | 0.13      | 0.33           |
| Age                             | -0.03    | -0.08, 0.03   | 0.33      | -0.08    | -0.17, -0.006 | 0.07      | 0.30           |
| Race/ethnicity                  |  |  |  |  |  |  |  |
| White, non-Hispanic (Reference) |  |  |  |  |  |  |  |
| Asian                           | -1.27    | -5.11, 2.56   | 0.52      | 0.91     | -3.99, 5.81   | 0.72      | 0.49           |
| Black, non-Hispanic             | -0.10    | -5.18, 4.98   | 0.97      | 6.47     | 0.82, 12.11   | 0.03      | 0.09           |
| Hispanic                       | 2.32     | -1.34, 5.98   | 0.21      | 3.40     | -1.09, 7.90   | 0.14      | 0.68           |
| Body mass index                 | 0.003    | -0.0001, 0.01 | 0.05      | -0.02    | -0.20, 0.16   | 0.84      | 0.85           |
| Smoking status                  | 4.05     | 0.43, 7.67    | 0.03      | 2.55     | -2.37, 7.47   | 0.31      | 0.60           |
| Biologic/small molecule DMARD   | -0.74    | -1.98, 0.49   | 0.24      | 1.56     | -2.06, 5.18   | 0.40      | 0.19           |
| Synthetic DMARD                 | -1.38    | -2.76, 0.002  | 0.05      | 1.25     | -0.98, 3.48   | 0.27      | 0.07           |
| Time (per year)                 | -2.11    | -4.43, -0.21  | 0.07      | -3.22    | -6.02, -0.41  | 0.03      | 0.51           |

Abbreviation: CI, confidence interval; DMARD, disease-modifying antirheumatic drug.

* Mixed effects models over 38-month period; interaction term for each variable by source of data (cohort vs. noncohort)

* Significant at $P$ interaction <0.10.
these factors at time of recruitment. This may be unique to the nature of the RA Observational Cohort study, which did not require separate visits beyond regular care or involve an intervention (eg, medication/therapy). Previous studies in non-RA populations have examined differences in patient outcomes between observational cohort and noncohort participants. Similar to our findings, Manjer et al found that although no differences were found with respect to sociodemographic variables at baseline, outcomes such as cancer incidence and mortality were higher in nonparticipants during and following recruitment (9). These findings indicate that although differences in outcome between participants and nonparticipants may not be evident at baseline, there could be substantial differences between groups over time, which may lead results to be unrepresentative of the greater patient population.

Our findings suggest that studies utilizing EHR records may provide different conclusions regarding disease activity trajectories than traditional cohort studies in some circumstances. We found that black, non-Hispanic patients had significantly higher disease activity as measured by CDAI compared with white, non-Hispanic patients amongst noncohort participants but not in cohort participants. Previous literature has suggested that certain racial/ethnic groups report higher disease activity with respect to self-reported race and ethnicity (6,10) and that differences persist over time (12), whereas other studies have shown no differences between groups (13). Lack of association found in cohort participants within our analyses indicate that unmeasured differences between groups at baseline may be associated with longitudinal disease outcomes. For example, the cohort group may consist of individuals likely to be more adherent to medications and/or consistently follow up on their care, given their participation in the study (ie, volunteer bias). This also may explain the differing associations of DMARDs on disease activity between the two groups.

Recent literature has highlighted that while EHR studies may capture a broader patient population, they still experience substantial challenges with missing data (1). EHR data are less standardized, and thus missing values of key variables may be significantly present. We found that compared with cohort participants, noncohort participants in the EHR had a higher percentage of missing data across a majority of variables. To ensure that all disease activity measures (including those mentioned in clinical notes and not entered in structured data) were included for analysis, we conducted chart reviews for each patient in the study. We found that chart review resulted in increased capture of 19% of CDAI scores, 4% of Patient Global scores, and 30% of Physician Global scores across the entire dataset of cohort and noncohort participants. Future efforts to standardize and make EHRs interoperable will help increase generalizability and completeness of EHR data (1), as will advances in natural language processing to extract this information in an accurate and reliable manner without the burden of chart review. Imputation of data across cohort and EHR data also may improve precision of estimates by increasing the number of observations and data values, since different types of information is available in the two datasets. We cannot rule out the possibility that the noncohort group represents a different phenotype of RA or misdiagnosis; however, we utilized an inclusion criteria that has demonstrated high positive predictive value of RA diagnosis and conducted sensitivity analyses that additionally required that patients be prescribed at least one DMARD during the study period. Lastly, we were unable to examine whether differences in disease activity trajectories were associated with certain social determinants of health, including poverty, homelessness, and stress, which are drastically underrepresented in EHRs but often collected in cohort studies (14,15).

This initial work suggests that data derived from the EHR uncovered important health disparities that were not observed in a cohort study. Future work should focus on addressing key limitations in use of EHR data, such as missingness, while also further exploring the potential strengths of this data source, such as the ability to uncover important health disparities.

**AUTHOR CONTRIBUTIONS**

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Gianfrancesco had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

**Study conception and design.** Gianfrancesco, McCulloch, Schmajuk, Yazdany.

**Acquisition of data.** Gianfrancesco, Trupin.

**Analysis and interpretation of data.** Gianfrancesco, Trupin, McCulloch, Shiboski, Schmajuk, Yazdany.

**REFERENCES**

1. Casey JA, Schwartz BS, Stewart WF, et al. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. Annu Rev Public Health 2016;37:61–81.
2. Damotte V, Lizee A, Tremblay M, et al. Harnessing electronic medical records to advance research on multiple sclerosis. Mult Scler 2018;1352458517747407.
3. Wallach JD, Sullivan PG, Trepanowski JF, et al. Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials. JAMA Intern Med 2017;177(4):554–60.
4. Ejogu N, Norbeck JH, Mason MA, et al. Recruitment and retention strategies for minority or poor clinical research participants: lessons from the Healthy Aging in Neighborhoods of Diversity across the Life Span study. Gerontologist 2011;51 Suppl 1:S33–45.
5. Falasiriu T, Chaichian Y, Bass MB, et al. 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.
6. 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):1238–46.
7. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31(3):315–24.
8. 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.

9. 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):489–99.

10. Yazici Y, Kautiainen H, Sokka T. Differences in clinical status measures in different ethnic/racial groups with early rheumatoid arthritis: implications for interpretation of clinical trial data. J Rheumatol 2007;34(2):311–5.

11. Bruce B, Fries JF, Murtagh KN. Health status disparities in ethnic minority patients with rheumatoid arthritis: a cross-sectional study. J Rheumatol 2007;34(7):1475–9.

12. Greenberg JD, Spruill TM, Shan Y, et al. Racial and ethnic disparities in disease activity in patients with rheumatoid arthritis. Am J Med 2013;126(12):1089–98.

13. Iren UT, Walker MS, Hochman E, et al. A pilot study to determine whether disability and disease activity are different in African-American and Caucasian patients with rheumatoid arthritis in St. Louis, Missouri, USA. J Rheumatol 2005;32(4):602–8.

14. Adler NE, Stead WW. Patients in context–EHR capture of social and behavioral determinants of health. N Engl J Med 2015;372(8):698–701.

15. Institute of Medicine (U.S.). Board on Population Health and Public Health Practice. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records. Capturing social and behavioral domains in electronic health records. Phase 1. Washington, D.C.: The National Academies Press 2014.