Concordance between survey and electronic health record data in the COVID-19 Citizen Science study: a retrospective cohort analysis

Elizabeth Crull, Emily C. O'Brien, Pavel Antiperovitch, Kirubel Asfaw, Alexis L. Beatty, Djeneba Audrey Djibo, Alan F. Kaul, John Kornak, Gregory M. Marcus, Madelaine Faulkner Modrow, Jeffrey E. Olgin, Jaime Orozco, Soo Park, Noah Peyser, Mark J. Pletcher, Thomas W. Carton

Submitted to: Journal of Medical Internet Research
on: March 05, 2024

Table of Contents

Concordance between survey and electronic health record data in the COVID-19 Citizen Science study: a retrospective cohort analysis

Elizabeth Crull[1] MPH; Emily C. O'Brien[2, 3] PhD; Pavel Antiperovitch[4] MD; Kirubel Asfaw[2] MS; Alexis L. Beatty[5] MD, MAS; Djeneba Audrey Djibo[6] PhD; Alan F. Kaul[7] PharmD, MBA; John Kornak[4] PhD; Gregory M. Marcus[5] MD, MAS; Madelaine Faulkner Modrow[4] MPH; Jeffrey E. Olgin[5] MD; Jaime Orozco[4] BA; Soo Park[4] BA; Noah Peyser[5] PhD; Mark J. Pletcher[4] MD, MPH; Thomas W. Carton[8] MS, PhD

[1]Louisiana Public Health Institute Department of Health Services Research New Orleans US
[2]Duke Clinical Research Institute School of Medicine Duke University Durham US
[3]Department of Epidemiology and Biostatistics University of California, San Francisco San Francisco US
[4]Division of Cardiology University of California, San Francisco San Francisco US
[5]Safety, Surveillance, and Collaboration CVS Health Blue Bell US
[6]Medical Outcomes Management, Inc. Sharon US
[7]Department of Health Services Research Louisiana Public Health Institute New Orleans US

Corresponding Author:
Elizabeth Crull MPH
Louisiana Public Health Institute
Department of Health Services Research
1515 Poydras Street
Suite 1250
New Orleans
US

Abstract

Background: Real-world data reported by patients and extracted from electronic health records is increasingly leveraged for research, policy, and clinical decision-making. However, it is not always obvious the extent to which these two data sources agree with each other.

Objective: To evaluate the concordance of variables reported by participants enrolled in an electronic cohort study and data available in their electronic health records.

Methods: Survey data from COVID-19 Citizen Science, an electronic cohort study, were linked to electronic health record data from 7 health systems, comprising 34,908 participants. Concordance was evaluated for demographics, chronic conditions, and COVID-19 characteristics. Overall agreement, sensitivity, specificity, positive predictive value, negative predictive value, and ? statistics with 95% CIs were calculated.

Results: Of 34,017 participants with complete information, 62.3% were female, and the median age was 57 (IQR, 42-68). Agreement (?) was high for sex (? = 0.99) and Black (? = 0.94), AAPI (? = 0.93), and White (? = 0.87) race and ethnicity but only moderate (? = 0.54) for smoking status. Compared with chart data, participant report of chronic conditions had lower sensitivity and higher specificity, with widely varying levels of agreement (?). Compared with participant report of COVID-19, electronic health record data had low sensitivity (32.2%) but higher specificity (95.8%). COVID-19 vaccination was the least concordant event (? = 0.05) but had moderate sensitivity (49.7%) and high sensitivity (98.2%) compared to participant reports.

Conclusions: Results suggest that additional work is required to integrate and prioritize participant-reported data in pragmatic research. Clinical Trial: ClinicalTrials.gov Identifier NCT5548803

(JMIR Preprints 05/03/2024:58097)
DOI: https://doi.org/10.2196/preprints.58097

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Authorship

Elizabeth Crull, MPH[1]; Emily C. O'Brien, PhD[2,3]; Pavel Antiperovitch, MD[4]; Kirubel Asfaw, MS[3]; Alexis L. Beatty, MD, MAS[5]; Djeneba Audrey Djibo, PhD[6]; Alan F. Kaul, PharmD, MBA[7]; John Kornak, PhD[4]; Gregory M. Marcus, MD, MAS[5]; Madelaine Faulkner Modrow, MPH[4]; Jeffrey E. Olgin, MD[5]; Jaime Orozco, BA[4]; Soo Park, BA[4]; Noah Peyser, PhD[5]; Mark J. Pletcher, MD, MPH[4]; Thomas W. Carton, PhD[1]

Author Affiliations

[1]Department of Health Services Research, Louisiana Public Health Institute, New Orleans, LA, USA [2]Department of Population Health Sciences, Duke University School of Medicine, Durham, North Carolina, USA [3]Duke Clinical Research Institute, Duke University School of Medicine, Durham, North Carolina, USA [4]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA [5]Division of Cardiology, Department of Medicine, University of California San Francisco, San Francisco, CA, USA [6]Safety, Surveillance & Collaboration, CVS Health, Blue Bell, PA, USA [7]Medical Outcomes Management, Inc., Sharon, MA, USA

Corresponding Author

Thomas W Carton, PhD, Department of Health Services Research, Louisiana Public Health Institute, 400 Poydras St., Suite 1250, New Orleans, LA 70130 E: tcarton@lphi.org P: 504-715-6726

Concordance between survey and electronic health record data in the COVID-19 Citizen Science study: a retrospective cohort analysis

Abstract

Background: Real-world data reported by patients and extracted from electronic health records is increasingly leveraged for research, policy, and clinical decision-making. However, it is not always obvious the extent to which these two data sources agree with each other.

Objective: To evaluate the concordance of variables reported by participants enrolled in an electronic cohort study and data available in their electronic health records.

Methods: Survey data from COVID-19 Citizen Science, an electronic cohort study, were linked to electronic health record data from 7 health systems, comprising 34,908 participants. Concordance was evaluated for demographics, chronic conditions, and COVID-19 characteristics. Overall agreement, sensitivity, specificity, positive predictive value, negative predictive value, and κ statistics with 95% CIs were calculated.

Results: Of 34,017 participants with complete information, 62.3% were female, and the median age was 57 (IQR, 42-68). Agreement (κ) was high for sex (κ = 0.99) and Black (κ = 0.94), AAPI (κ = 0.93), and White (κ = 0.87) race and ethnicity but only moderate (κ = 0.54) for smoking status. Compared with chart data, participant report of chronic conditions had lower sensitivity and higher specificity, with widely varying levels of agreement (κ). Compared with participant report of COVID-19, electronic health record data had low sensitivity (32.2%) but higher specificity (95.8%). COVID-19 vaccination was the least concordant event (κ = 0.05) but had moderate sensitivity (49.7%) and high sensitivity (98.2%) compared to participant reports.

Conclusion: Results suggest that additional work is required to integrate and prioritize participant-reported data in pragmatic research.

Trial Registration: ClinicalTrials.gov Identifier NCT5548803

Keywords: Electronic health records; self-report; COVID-19; data accuracy; data validation

Introduction

The advent of electronic health record (EHR) systems and internet-based study portals have modernized and streamlined pragmatic clinical research,[1-2] defined as research that can be conducted in real-world settings with minimal change to clinical operation.[3-4] The use of patient- and study participant-reported data alongside EHR data is increasingly common in research and clinical practice to complement and validate EHR data sources.[5-6] Therefore, there is potential value in linking and comparing patient experience and outcomes data gathered from mailed surveys[7-9] and patient-facing online portals[10] with data extracted from EHRs. This is especially true for clinical concepts that are notoriously difficult to qualify using medical coding alone like mood, gastrointestinal disorders, and chronic pain.[11]

EHR and participant-reported data each have significant limitations. EHR data are fraught with administrative error, incomplete mapping to clinical ontologies, lack of legacy health record data, and inability to extract important clinical information from unstructured physician notes.[12-13] Participant-reported data are subject to bias from social desirability,[10] fatigue,[14] and limited understanding of medical issues. How these limitations affect the reliability of different kinds of health-related information is unclear. There is a particular need for understanding the reliability of health information related to COVID-19, especially because much home testing and vaccination occurred outside traditional health systems.

The COVID-19 Citizen Science Study (CCS) is a longitudinal digital cohort study designed to generate knowledge about participant-reported outcomes related to the COVID-19 pandemic.[15] The study linked participant reported data with their corresponding EHR data, thus presenting opportunity to analyze the concordance between these data sources. We analyzed these data to assess concordance of COVID-19-related outcomes, demographic characteristics, smoker status, and 12 common medical conditions.

Methods

Study Recruitment

Our concordance assessment used participant-reported and EHR data collected as part of the COVID-19 Citizen Science (CCS) study (ClinicalTrials.gov Identifier NCT5548803), which has been described in detail previously.[15] Participants were recruited from 7 major health systems in Texas, Louisiana, Mississippi, California, Utah, and New York that participate in the National Patient-Centered Clinical Research Network (PCORnet). Patients were eligible to join if they were aged 18 years or older and had at least 1 clinical encounter after January 1, 2019. Recruitment lasted from November 2020 to February 2022.

Participant-Reported Data

Upon enrolling, participants were asked to respond to baseline surveys on demographics, smoking history, and medical conditions. Participants were then administered follow-up surveys about exposure to, diagnosis of, and vaccination against COVID-19, among other questions seeking to understand both individual experience and population-level trends related to the pandemic. These surveys were housed in the Eureka research platform (University of California San Francisco with funding from National Institutes of Health),[16] which had web browser and smartphone functionality.

Electronic Health Record Data

For consenting and authorizing participants, EHR limited dataset in the PCORnet Common Data Model (CDM) format were extracted from the site-specific DataMarts maintained by all participating health systems.[17] The CCS study data extraction query was developed by Duke University programmers using SAS® 9.4 software and distributed to all sites to run in their local environments against their DataMart. The query extracted clinical data with a 5-year lookback from recruitment

start date through the most recently available data. Sensitive diagnoses were filtered out, and only a minimum necessary subset of laboratory and medication records were extracted. Only patients for whom identities were algorithmically matched or manually verified were included in the final analytic cohort.

Concordance Definitions

Among 34,908 participants where linkage was possible, we evaluated concordance in the following domains: demographics, baseline medical conditions, current smoker status, COVID-19 diagnosis, and COVID-19 vaccination. We chose variables that were conceptually similar between the participant-reported and EHR sources (**Supplement 1, 2**). Sex in both sources was defined as sex assigned at birth. Although gender identity was available from survey data, it was not available in EHR data and thus not an eligible variable for concordance analysis. Race and ethnicity data abstracted from EHR data were populated according to health system practices. Race and ethnicity data abstracted from survey data were reported directly by study participants.

To promote comparability, measurement periods were aligned between data sources. Participants who had missing data in one or both sources were not considered for concordance analyses. Age data were not analyzed for concordance because a birthdate match between sources was a requirement for data to be considered for EHR data extraction; thus, discordant scenarios were inherently filtered out prior to analysis for this substudy.

For demographic, smoker status, and COVID-19 characteristics, the participant report was considered the criterion standard. For medical conditions, the EHR was considered the criterion standard.

Statistical Approach

To test for marginal homogeneity between data sources, the McNemar test for paired nominal data was run on each characteristic set up as a dichotomous 2x2 contingency table. Chi-square (χ^2)

statistics and p-values were generated, and p-values less than .001 were reported as p<.001.

For all domains, the following statistics were generated along with their 95% confidence intervals: overall agreement, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Cohen's kappa (κ). We used the following ranges for κ to describe observed agreement: strong (0.81-1.00), good (0.61-0.80), moderate (0.41-0.60), fair (0.21-0.40), poor (0.01-0.20), and no agreement (<0).[18-19] However, these ranges are to provide a guide, and the adequacy of the specific level of agreement should be considered specifically to the domain under consideration and the application to which it will be used.

Data were analyzed from December 2022 to July 2023 using SAS® 9.4 software. The CCS study, the protocol for which covered data analysis activities conducted for this substudy, was approved by Western IRB on November 5, 2020.

Results

39,526 patients enrolled in the study across 7 sites. After exclusion of participants whose identity could not be verified and participants with partially or completely missing EHR data, 34,908 participants were included in the final analytic cohort (**Figure 1**). Descriptive statistics of the 34,017 participants who responded to the baseline demographic survey and results from the McNemar test for marginal homogeneity between data sources are summarized in **Table 1**. The median age of the sample was 57 years old (range 18-100, interquartile range 42-68) according to the EHR. The sample was predominantly female and white according to both sources. The sample was classified as 6.6% Hispanic in the EHR and 9.2% Hispanic according to participant-reported data (p<.001).

Figure 1. Enrollment Diagram and Final Analytic Cohort for CCS Concordance Substudy

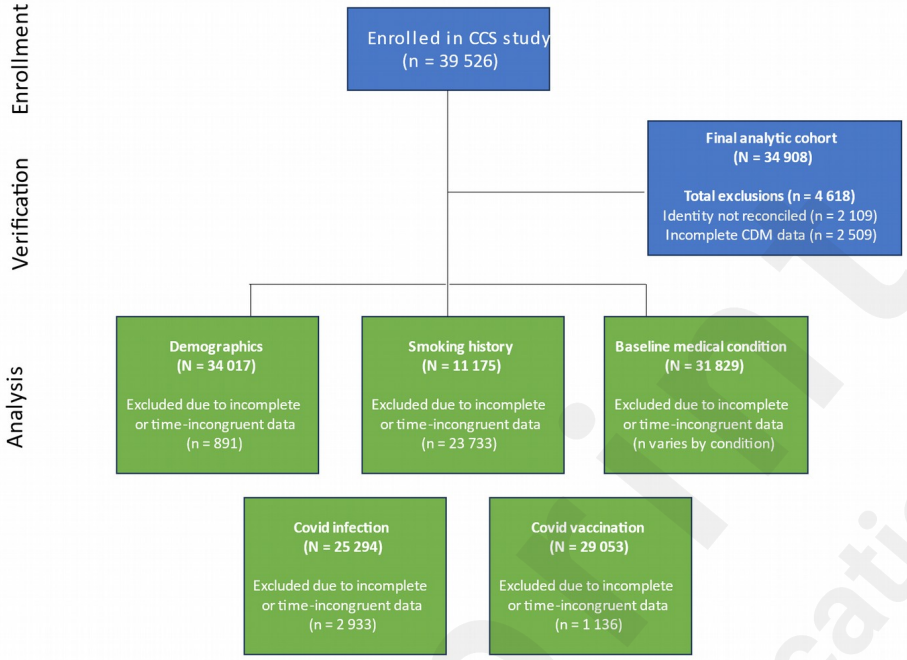

Table 1. EHR- and Participant Report-Derived Demographic Information

Variable	EHR	Participant Report	P Value
	No. (%)	**No. (%)**	**McNemar Test**
N	**34,017**	**34,017**	
Age, years[a]			
Mean (SD)	54.7 (16.1)		
Median (IQR) [range]	57 (42-68) [18-100]		
Sex			
Female	21217 (62.4)	21176 (62.3)	.005
Male	12780 (37.6)	12742 (37.5)	.007
Refused or missing	20 (<1)	99 (<1)	<.001
Race			
American Indian or Alaska Native	104 (<1)	91 (<1)	.21
Asian, Native Hawaiian, or Pacific Islander	1797 (5.3)	2042 (6.0)	<.001
Black or African American	1286 (3.8)	1344 (4.0)	.001
White	27054 (79.5)	27744 (81.6)	<.001
Multiple Races[b]	93 (<1)	1269 (3.7)	<.001
Other[c]	1039 (3.1)	1077 (3.2)	.33
Refused or missing	2644 (7.8)	450 (1.3)	<.001
Ethnicity			
Hispanic	2249 (6.6)	3124 (9.2)	<.001
Non-Hispanic	28903 (85.0)	30528 (89.7)	<.001
Refused or missing	2865 (8.4)	365 (1.1)	<.001

[a] Age data were not analyzed for concordance because a DOB match between sources was a requirement for data to be considered for EHR data extraction; thus discordant scenarios were inherently filtered out prior to analysis for this substudy

[b]The category “Multiple Races” is a mapped value in the PCORnet Common Data Model and no further detail was available. In participant-reported data, “Multiple Races” was defined as participants who responded to 2 or more non-missing race categories.

[c] In both data sources, there were no additional details for the “Other” categorization.

Statistically significant differences between the two data sources were detected for all characteristics except for diabetes (p=.17). The starkest absolute difference between data sources was for COVID-19 vaccination, in which 97.4% of participants self-reported a vaccine while only 48.4% had this documented in the EHR (**Table 2**).

Table 2. EHR- and Participant Report-Derived Medical Condition and Covid-19 Information

Variable	N	EHR	Participant Report	P Value
		%	%	McNemar Test
Smoking status	11175	8.5	12.5	<.001
Medical conditions				
Diabetes	31744	9.6	9.4	.17
Hypertension	31636	29.8	34.6	<.001
Coronary artery disease/angina	31573	7.7	6.1	<.001
Myocardial infarction	31721	0.9	2.5	<.001
Congestive heart failure	31686	2.2	1.7	<.001
Transient ischemic attack	31659	1.7	3.0	<.001
Atrial fibrillation/flutter	31495	4.3	5.6	<.001
Sleep apnea	31075	9.1	15.8	<.001
COPD	31626	4.6	3.5	<.001
Asthma	31738	11.0	9.9	<.001
Immunodeficiency	31378	2.3	5.1	<.001
Anemia	31622	12.1	10.8	<.001
Covid-19				
Infection	25294	9.2	17.6	<.001
Vaccination	29053	48.4	97.4	<.001

Agreement between EHR and participant-reported characteristics according to 5 proportionate measures (overall agreement, sensitivity, specificity, PPV, NPV) and 1 statistic of interrater reliability (κ) are shown in **Table 3, Supplement 3**. Overall agreement was above 95.0% for all demographic characteristics, where the participant report was considered the criterion standard. Chance-corrected agreement (κ) was strong for most demographic characteristics. Sensitivity was 74.0% for the Hispanic characteristic, which translates to a relatively higher number of false negatives compared to other racial groups.

Table 3. Agreement for EHR and Participant-Reported Characteristics

	% (95% CI)					
Variable	**Overall agreement**	**κ Statistic**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
Female	99.6 (99.5-99.7)	0.99 (0.99-0.99)	99.6 (99.5-99.7)	99.4 (99.3-99.6)	99.7 (99.6-99.7)	99.4 (99.2-99.5)
Non-Hispanic						
AAPI	99.2 (99.1-99.3)	0.93 (0.92-0.94)	92.0 (90.7-93.3)	99.6 (99.6-99.7)	93.9 (92.8-95.1)	99.5 (99.4-99.6)
Black	99.6 (99.5-99.7)	0.94 (0.93-0.95)	97.2 (96.2-98.2)	99.7 (99.6-99.7)	91.6 (90.0-93.2)	99.9 (99.9-99.9)
White	96.1 (95.9-99.3)	0.87 (0.86-0.88)	99.2 (99.0-99.3)	83.3 (82.3-84.3)	96.2 (95.9-96.4)	95.9 (95.3-96.4)
Other	96.9 (96.7-97.1)	0.31 (0.27-0.36)	26.6 (23.6-29.6)	99.0 (98.8-99.1)	42.7 (38.4-47.0)	97.9 (97.7-98.1)
Hispanic	97.9 (97.7-98.0)	0.82 (0.81-0.83)	74.0 (72.1-75.9)	99.7 (99.6-99.8)	94.9 (93.8-96.0)	98.1 (97.9-98.2)
Current smoker	91.4 (90.9-91.9)	0.54 (0.52-0.57)	49.4 (46.8-52.1)	97.4 (97.1-97.7)	72.9 (70.0-75.7)	93.1 (92.6-93.6)
Diabetes	95.1 (94.9-95.3)	0.71 (0.70-0.73)	73.5 (71.9-75.1)	97.4 (97.2-97.6)	74.8 (73.3-76.4)	97.2 (97.0-97.4)
Hypertension	85.0 (84.6-85.4)	0.66 (0.64-0.68)	82.9 (82.1-83.6)	85.9 (85.4-86.4)	71.4 (70.6-72.3)	92.2 (91.8-92.6)
Coronary artery disease/angina	94.0 (93.7-94.3)	0.54 (0.52-0.56)	51.0 (49.0-53.0)	97.6 (97.4-97.8)	64.1 (62.0-66.2)	96.0 (95.7-96.2)
Myocardial infarction	97.6 (97.5-97.8)	0.30 (0.25-0.35)	57.6 (51.9-63.3)	98.0 (97.9-98.2)	21.0 (18.1-23.8)	99.6 (99.5-99.7)
Congestive heart failure	98.0 (97.9-98.2)	0.48 (0.44-0.52)	44.1 (40.4-47.8)	99.2 (99.1-99.3)	55.8 (51.6-60.0)	98.8 (98.6-98.9)
Transient ischemic attack	97.6 (97.4-97.7)	0.47 (0.43-0.50)	66.2 (62.2-70.2)	98.1 (97.9-98.3)	37.4 (34.3-40.4)	99.4 (99.3-99.5)
Atrial fibrillation	97.0 (96.8-97.2)	0.68 (0.66-0.70)	80.3 (78.2-82.4)	97.8 (97.6-98.0)	61.8 (59.6-64.1)	99.1 (99.0-99.2)
Sleep apnea	90.4 (90.0-90.7)	0.56 (0.55-0.58)	83.5 (82.2-84.9)	91.0 (90.7-91.4)	48.4 (47.0-49.8)	98.2 (98.1-98.4)
COPD	95.2 (95.0-95.5)	0.39 (0.36-0.42)	36.7 (34.2-39.2)	98.1 (97.9-98.3)	48.4 (45.5-51.4)	97.0 (96.8-97.1)
Asthma	90.2 (89.9-90.6)	0.48 (0.46-0.50)	50.8 (49.1-52.5)	95.1 (94.9-95.4)	56.1 (54.4-57.9)	94.0 (93.7-94.3)
Immunodeficiency	94.6 (94.4-94.9)	0.26 (0.22-0.29)	45.1 (41.5-48.7)	95.8 (95.6-96.0)	20.4 (18.5-22.4)	98.7 (98.5-98.8)
Anemia	85.0 (84.6-85.4)	0.26 (0.24-0.28)	32.8 (31.3-34.3)	92.2 (91.9-92.5)	36.7 (35.1-38.3)	90.8 (90.5-91.2)
COVID-19 infection	84.6 (84.1-85.0)	0.34 (0.33-0.36)	32.2 (30.9-33.6)	95.8 (95.5-96.0)	61.8 (59.8-63.8)	86.9 (86.5-87.3)
COVID-19 vaccination	51.0 (50.4-51.6)	0.05 (0.04-0.06)	49.7 (49.1-50.3)	98.2 (97.2-99.1)	99.9 (99.9-99.9)	5.0 (4.6-5.3)

The criterion standard for baseline medical conditions was the EHR. Overall agreement, specificity, and NPV between data sources was above 85.0% for all baseline medical conditions, though there was heterogeneity in sensitivity, PPV, and chance-corrected agreement (κ). Sensitivity ranged from 32.8% [31.3-34.3] to 83.5% [82.2-84.9], being lowest for anemia and highest for apnea. PPV ranged from 20.4% [18.4-22.4] to 74.8% [73.3-76.4], being lowest for immunodeficiency and highest for diabetes. Finally, chance-corrected agreement (κ) was good for 3 of 12 compared baseline medical conditions, moderate for 5, and fair for 4. Chance-corrected agreement (κ) ranged from 0.26 [0.22-0.29] to 0.71 [0.70-0.73] for immunodeficiency and diabetes, respectively.

The criterion standard for COVID-19 variables was participant-reported data. While chance-corrected agreement was fair for COVID-19 infection (0.34), it was poor for COVID-19 vaccination (0.05). Of 25,294 participants whose COVID-19 infection data could be compared, there were 3,899 cases of discordance, 77.3% of which were classified as the participant reporting a COVID-19 diagnosis but this not being reflected in the EHR. Similarly, of the 29,053 participants whose COVID-19 vaccine data could be compared, there were 14,243 cases of discordance, 99.9% of which were classified as the participant reporting a COVID-19 vaccine but this not being reflected in the EHR.

Discussion

Principal Results

We evaluated the concordance survey data from participants enrolled in a COVID-19 study and their linked EHR data. We had 3 main findings: 1) agreement between the 2 sources was high for sex, race, and ethnicity but only moderate for smoking status; 2) agreement for baseline

comorbidities was lower and varied widely by condition, with the highest agreement observed for diabetes; and 3) chance-corrected agreement was fair to poor for COVID-19 infection and vaccination.

Increasingly, novel research designs rely on integration of multiple data sources to answer research questions. The COVID-19 pandemic accelerated already growing interest in real-world data use cases,[20] including leveraging existing EHR data and bringing research directly to people through participant-facing portals. Direct-to-participant research has numerous benefits, including potential for greater geographic reach and diversity, lower participant burden with few or no in-person visits, and platforms that enable capture of relevant patient-centered endpoints.[21] These strengths complement those of EHR data, which, through national networks like PCORnet, can be standardized into research-grade data to facilitate rapid insights into key clinical outcomes.[22] To our knowledge, this is the first study to examine patterns of concordance across participant-reported and EHR data in the context of COVID-19.

Our finding that participants self-reported COVID-19 infection and vaccination at higher rates than what was evident in their clinical records illustrates the fragmented nature of real-world data. Ongoing work to enhance the quality and reliability of EHR data in the context of COVID-19, including network-level curation,[22] linkage to external sources where appropriate (e.g., state vaccine[23] or policy[24] databases), and systematic phenotype development and testing[25] are critical to maximizing the research value of these data. In parallel, implementation of best-practices to enhance validity of participant-report, including stakeholder engagement in survey design, readability assessments, and cultural and linguistic adaptation are essential to enhancing reliability of findings from participant-facing research.[26-27] Our findings are broadly consistent with those from prior studies suggesting that fitness-for-use

depends on context.[28-29] We found that no single data source may be appropriate for EHR-based pragmatic research, consistent with prior work illustrating the potential biases that can arise in participant-reported data and how they vary.[30-33]

Limitations

Several limitations to our study are worth noting. First, the CCS study comprises participants that were mostly white and female. Therefore, results may not generalize to broader populations. Second, diversity within minority communities can make both responding to survey questions and the identification of race challenging for patients and clinicians, respectively, a fact that may skew findings from our demographic analyses. Third, participant-reported COVID-19 variables were not validated and are subject to reporting and recall bias. Fourth, we observed some attrition in reporting over time, which could lead to selection bias in analyses of longitudinal outcomes. Finally, we used EHR data for this concordance analysis, which may not represent all medical encounters for a given participant and which may not be of the highest or most accurate quality. EHR data used in this analysis were not linked to claims data from pharmacies, which are a major administrator of COVID-19 vaccines. Particularly for outcomes that are generally observed outside of the hospital, linkage to external data sources is likely warranted.

Conclusions

We found that integration of multiple data sources to investigate COVID-19 research questions enhance capture of key elements but also introduce opportunities for disagreement. Future studies that leverage linked data should evaluate concordance of overlapping elements and report levels of agreement. Transparent reporting will contribute to broader understanding of data reliability and relevance and support future strategies to improve fitness-for-use of real-world

data.

Acknowledgments

Author Contributions: EC had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

- *Concept and design:* Carton, Crull, O'Brien, Pletcher.
- *Acquisition, analysis, or interpretation of data:* Carton, Crull, O'Brien, Pletcher.
- *Drafting of the manuscript:* Carton, Crull, O'Brien.
- *Critical revision of the manuscript for important intellectual content:* All authors.
- *Statistical analysis:* Crull.
- *Obtained funding:* Carton, Pletcher.
- *Administrative, technical, or material support:* Asfaw, Modrow, Orozco, Park.
- *Supervision:* Carton, O'Brien, Pletcher.

Data Sharing Statement: Aggregated 2x2 contingency tables used to run agreement analyses are available in Supplement 3. The full code list and SAS program used for the electronic health record data queries are available upon request from the first author, EC. Survey questions/responses are available upon request from the first author, EC. In an effort to protect patient privacy, individual-level data are subject to privacy/ethical restrictions and are not publicly available.

Competing Interests: None declared

Funding/Support: This work was supported with funding from the Patient Centered Outcomes Research Institute (PCORI). Grant Identification: COVID-2020C2-10761

Role of the Funder/Sponsor: PCORI had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval

of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The views and conclusions presented here are solely the responsibility of the authors and do not necessarily reflect the official views of PCORI.

Additional Information: The full code list and SAS program used for the electronic health record data queries are available upon request. Survey questions/responses are available upon request.

References

1. Cowie, M. R., Blomster, J. I., Curtis, L. H., et al (2017). Electronic health records to facilitate clinical research. Clinical research in cardiology : official journal of the German Cardiac Society, 106(1), 1–9. https://doi.org/10.1007/s00392-016-1025-6

2. van Staa, T.-P., Goldacre, B., Gulliford, M., et al. (2012). Pragmatic randomised trials using routine electronic health records. BMJ: British Medical Journal, 344(7843), 22–25. http://www.jstor.org/stable/41502081

3. Tosh, G., Soares-Weiser, K., & Adams, C. E. (2011). Pragmatic vs explanatory trials: the pragmascope tool to help measure differences in protocols of mental health randomized controlled trials. Dialogues in clinical neuroscience, 13(2), 209–215. https://doi.org/10.31887/DCNS.2011.13.2/gtosh

4. Holtrop, J. S., & Glasgow, R. E. (2020). Pragmatic research: an introduction for clinical practitioners. Family Practice, 37(3), 424-428. https://doi.org/10.1093/fampra/cmz092

5. Black, N. (2013). Patient reported outcome measures may transform healthcare. BMJ: British Medical Journal, 346(7896), 19–21. http://www.jstor.org/stable/23494165

6. Jandoo T. (2020). WHO guidance for digital health: What it means for researchers. Digital health, 6, 2055207619898984. https://doi.org/10.1177/2055207619898984

7. Hamilton, N. S., Edelman, D., Weinberger, M., et al (2009). Concordance between self-reported race/ethnicity and that recorded in a Veteran Affairs electronic medical record. North Carolina medical journal, 70(4), 296–300.

8. Valikodath, N. G., Newman-Casey, P. A., Lee, P. P et al (2017). Agreement of Ocular Symptom Reporting Between Patient-Reported Outcomes and Medical Records. JAMA ophthalmology, 135(3), 225–231. https://doi.org/10.1001/jamaophthalmol.2016.5551

9. Fares, C. M., Williamson, T. J., Theisen, M. K., et al (2018). Low Concordance of Patient-Reported Outcomes With Clinical and Clinical Trial Documentation. JCO clinical cancer informatics, 2, 1–12. https://doi.org/10.1200/CCI.18.00059

10. O'Brien, E. C., Mulder, H., Jones, W. S., et al (2022). Concordance Between Patient-Reported Health Data and Electronic Health Data in the ADAPTABLE Trial. JAMA cardiology, 7(12), 1235–1243. https://doi.org/10.1001/jamacardio.2022.3844

11. Hostetter, M. & Klein, S. (n.d.). Using Patient-Reported Outcomes to Improve Health Care Quality. The Commonwealth Fund. https://www.commonwealthfund.org/publications/newsletter-article/using-patient-reported-outcomes-improve-health-care-quality

12. Verheij, R. A., Curcin, V., Delaney, B. C., et al (2018). Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. Journal of medical Internet research, 20(5), e185. https://doi.org/10.2196/jmir.9134

13. Menachemi, N., & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. Risk management and healthcare policy, 4, 47–55. https://doi.org/10.2147/RMHP.S12985

14. Zini, M. L. L., & Banfi, G. (2021). A Narrative Literature Review of Bias in Collecting Patient Reported Outcomes Measures (PROMs). International journal of environmental research

and public health, 18(23), 12445. https://doi.org/10.3390/ijerph182312445

15. Beatty, AL, Peyser, ND, Butcher, XE, et al (2021). The COVID-19 Citizen Science Study: Protocol for a Longitudinal Digital Health Cohort Study. JMIR research protocols, 10(8), e28169. https://doi.org/10.2196/28169

16. Peyser ND, Marcus GM, Beatty AL, et al. Digital platforms for clinical trials: The Eureka experience. Contemp Clin Trials. 2022 Apr;115:106710. doi: 10.1016/j.cct.2022.106710. Epub 2022 Feb 17. PMID: 35183763.

17. Forrest CB, McTigue KM, Hernandez AF, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol. 2021 Jan;129:60-67. doi: 10.1016/j.jclinepi.2020.09.036. Epub 2020 Sep 28. PMID: 33002635; PMCID: PMC7521354.

18. Altman, Douglas G. 1999. Practical Statistics for Medical Research. Chapman; Hall/CRC Press.

19. McHugh M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276–282.

20. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. JAMA 2018; 320(9):867-868.

21. de Jong AJ, van Rijssel TI, Zuidgeest MGP, et al. Opportunities and Challenges for Decentralized Clinical Trials: European Regulators' Perspective. Clin Pharmacol Ther 2022; 112(2):344-352.

22. Qualls LG, Phillips TA, Hammill BG, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet(R)). EGEMS (Wash DC) 2018; 6(1):3.

23. Groom HC, Crane B, Naleway AL, et al. Monitoring vaccine safety using the vaccine safety

Datalink: Assessing capacity to integrate data from Immunization Information systems. Vaccine 2022; 40(5):752-756.

24. Hamad R, Lyman KA, Lin F, et al. The U.S. COVID-19 County Policy Database: a novel resource to support pandemic-related research. BMC Public Health 2022; 22(1):1882.

25. Lusczek ER, Ingraham NE, Karam BS, et al. Characterizing COVID-19 clinical phenotypes and associated comorbidities and complication profiles. PLoS One 2021; 16(3):e0248956.

26. Chang EM, Gillespie EF, Shaverdian N. Truthfulness in patient-reported outcomes: factors affecting patients' responses and impact on data quality. Patient Relat Outcome Meas 2019; 10:171-186.

27. Breeman S, Constable L, Duncan A, et al. Verifying participant-reported clinical outcomes: challenges and implications. Trials 2020; 21(1):241.

28. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products. September 2021. Accessed December 22, 2021. Available at: https://www.fda.gov/media/152503/download. Accessed March 30, 2022.

29. Considerations for the Use of Real-World Data and Real-World Evidence to Support Regulatory Decision-Making for Drug and Biological Products. U.S. Department of Health and Human Services Food and Drug Administration. December 2021. Accessed December 22, 2021. Available at: https://www.fda.gov/media/154714/download. Accessed March 30, 2022.

30. Heckbert SR, Kooperberg C, Safford MM, et al. Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the Women's Health Initiative. Am J Epidemiol 2004; 160(12):1152-1158.

31. Stirratt MJ, Dunbar-Jacob J, Crane HM, et al. Self-report measures of medication adherence

behavior: recommendations on optimal use. Transl Behav Med 2015; 5(4):470-482.

32. Woodfield R, Group UKBSO, Follow-up UKB, Outcomes Working G, Sudlow CL. Accuracy of Patient Self-Report of Stroke: A Systematic Review from the UK Biobank Stroke Outcomes Group. PLoS One 2015; 10(9):e0137538.

33. Simpson CF, Boyd CM, Carlson MC, et al. Agreement between self-report of disease diagnoses and medical record validation in disabled older women: factors that modify agreement. J Am Geriatr Soc 2004; 52(1):123-127.

Supplementary Files

Untitled.

Figure 1. Enrollment Diagram and Final Analytic Cohort for CCS Concordance Substudy