Comparison of family health history in surveys vs electronic health record data mapped to the observational medical outcomes partnership data model in the All of Us Research Program

Robert M. Cronin
Alese E. Halvorson
Cassie Springer
Xiaoke Feng
Lina Sulieman

See next page for additional authors

Follow this and additional works at: https://scholarlycommons.henryford.com/chphsr_articles
Authors
Robert M. Cronin, Alese E. Halvorson, Cassie Springer, Xiaoke Feng, Lina Sulieman, Roxana Loperena-Cortes, Kelsey Mayo, Robert J. Carroll, Qingxia Chen, Brian K. Ahmedani, Jason Karnes, Bruce Korf, Christopher J O'Donnell, Jun Qian, and Andrea H. Ramirez
Comparison of family health history in surveys vs electronic health record data mapped to the observational medical outcomes partnership data model in the All of Us Research Program

Robert M. Cronin 1,2, Alese E. Halvorson,1 Cassie Springer,1 Xiaoke Feng,1 Lina Sulieman,1 Roxana Loperena-Cortes,1 Kelsey Mayo,1 Robert J. Carroll,1 Qingxia Chen,1 Brian K. Ahmedani,3 Jason Karnes,4 Bruce Korf,5 Christopher J. O’Donnell,6,7 Jun Qian,3 and Andrea H. Ramirez,1 All of Us Research Program Demonstration Projects Subcommittee, on behalf of the All of Us Research Program Investigators

1Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, 2Department of Medicine, The Ohio State University, Columbus, Ohio, USA, 3Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, Michigan, USA, 4Department of Pharmacy Practice and Science, University of Arizona College of Pharmacy, Tucson, Arizona, USA, 5Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama, USA, 6Department of Medicine, Veterans Administration Boston Healthcare System, Boston, Massachusetts, USA and 7Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, USA

Corresponding Author: Robert M. Cronin, MD, MS, 2525 West End, Suite 1475, Nashville, TN 37203, USA; robert.cronin@osumc.edu

Received 10 August 2020; Revised 15 October 2020; Editorial Decision 29 November 2020; Accepted 14 November 2020

ABSTRACT

Objective: Family health history is important to clinical care and precision medicine. Prior studies show gaps in data collected from patient surveys and electronic health records (EHRs). The All of Us Research Program collects family history from participants via surveys and EHRs. This Demonstration Project aims to evaluate availability of family health history information within the publicly available data from All of Us and to characterize the data from both sources.

Materials and Methods: Surveys were completed by participants on an electronic portal. EHR data was mapped to the Observational Medical Outcomes Partnership data model. We used descriptive statistics to perform exploratory analysis of the data, including evaluating a list of medically actionable genetic disorders. We performed a subanalysis on participants who had both survey and EHR data.

Results: There were 54 872 participants with family history data. Of those, 26% had EHR data only, 63% had survey only, and 10.5% had data from both sources. There were 35 217 participants with reported family history of a medically actionable genetic disorder (9% from EHR only, 89% from surveys, and 2% from both). In the subanalysis, we found inconsistencies between the surveys and EHRs. More details came from surveys. When both mentioned a similar disease, the source of truth was unclear.

Conclusions: Compiling data from both surveys and EHR can provide a more comprehensive source for family health history, but informatics challenges and opportunities exist. Access to more complete understanding of a person’s family health history may provide opportunities for precision medicine.

Key words: Family health history, precision medicine, health surveys, electronic health records

© The Author(s) 2021. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved. For permissions, please email: journals.permissions@oup.com
INTRODUCTION

Family health history is long recognized to be important in clinical care and precision medicine.1–3 Through family history, healthcare providers can recognize the potential risk of disease in an individual based on inheritance. By understanding an individual’s disease risk, early interventions can help mitigate or prevent familial disease occurrence. Harnessing informatics tools like predictive modeling and machine learning models can significantly improve the precise care that we deliver to patients.

Prior studies show gaps in data from electronic health record (EHR) structured fields that are moderately assisted by free text extraction; however, there are significant limitations in routine acquisition of family health history from EHRs.4–6 Family health history can also be collected through patient surveys. Patient surveys or questionnaires show promise in obtaining family history.5,7,8 However, similar to EHRs, patient surveys may collect limited data.7 While both EHRs and patient surveys could have more complete family health history, single-center studies combining data sources demonstrate limitations.9,10 There is a paucity of literature describing a publicly available using both data sources to describe family history of a large diverse cohort across the United States.

The All of Us Research Program (All of Us) is recruiting 1 million or more participants reflecting the rich diversity of the U.S. population to advance the science of precision medicine (https://www.researchallofus.org/).11 All of Us is collecting unique family health history information using online surveys and EHR data from participants recruited at multiple healthcare organizations across the country. Both data types are mapped to the Observational Medical Outcomes Partnership (OMOP) common data model (https://github.com/OHDSI/CommonDataModel/wiki). These data are then curated with algorithms to develop a clean dataset, followed by privacy rules to ensure participant de-identification. Additionally, fields containing identifiable information are removed. Curated data are deposited into the All of Us Researcher Workbench, a cloud-based analytic platform for researchers which includes custom graphical interface data selection tools and client support for Python or R in Jupyter notebooks. All of Us developed a set of Demonstration Projects to highlight the ability of the initial launch data and tools to answer pertinent research questions as well as address limitations of the data.12 This Demonstration Project aims to evaluate the availability of family health history information within the All of Us registered tier data and to characterize the structured data elements from both data sources.

MATERIALS AND METHODS

We analyzed data from the September 2019 All of Us registered tier version R2019Q4R3. This dataset included family health history surveys administered to participants between May 2018 and September 30, 2019, as well as EHR data between 1980 and 2019. For the purposes of this project, an observation is the mention of the presence or absence of a family history of a certain disease. Each participant can have multiple family history observations, as they may have more than 1 family health history condition in more than 1 relative. We analyzed data at both the participant and observation levels.

The All of Us Research Program Institutional Review Board has established that registered tier data available on the Researcher Workbench (https://workbench.researchallofus.org/) meet criteria for non–human subjects research. Therefore, this demonstration project did not require Institutional Review Board review.

Surveys

All of Us collects survey data through a participant portal which can be accessed through the Internet on a desktop computer, or via a downloadable app on a tablet or smartphone. The development and launch of these surveys are described elsewhere.13 There are 6 surveys: the first 3 are available when a participant enrolls, and the remaining surveys, including family health history, become available 90 days after initial enrollment.

Participants are asked questions about the health history of first-degree blood relatives, including mother, father, siblings, and children, as well as grandparents (Supplementary Appendix 1). Each response to a question that refers to a disease is considered an independent observation of family health history from the survey. Diseases are grouped by categories that cover most organ systems, including heart, lung, gastrointestinal, and endocrine. A participant may skip questions or respond “prefer not to answer” or “don’t know,” as well as “no blood-related siblings” and “no blood-related children.” The full survey is located online (https://www.researchallofus.org/data-sources/survey-explorer/). We excluded survey participants who indicated that they did not know any family health history and those who skipped every question. The family health history survey was designed to focus on medically actionable genetic disorders, guided by the list of genes/disorders recommended by the American College of Medical Genetics and Genomics for reporting secondary findings from genome sequencing.14 Therefore, all diseases have a corresponding question for each relative and can be mapped from the survey to the disease in a structured manner (ie, a participant can select any of the diseases in the list for each relative in the family health history survey).

Electronic health records

EHR data about family health history were mapped to OMOP using Systematized Nomenclature of Medicine (SNOMED) and International Classification of Diseases codes. These vocabularies contain medical condition codes for family history as well as codes for relatives with a family history; however, there are no singular codes that link a medical condition code to a specific relative. Data from each healthcare organization’s EHR are sent to a data repository at the Data Research Center for All of Us.

In this project, we identified family health history information in EHR data by reviewing the OMOP hierarchy for records with “family+history” or “FH:” anywhere in their OMOP concept name. We subsequently looked at parent concepts to identify any additional concepts that could be included. There were 235 unique OMOP concept names that satisfied these criteria (Supplementary Appendix 2). We excluded observations of “family social history” because family social history is related to a social situation like the death or absence of a family member and is not considered a medical condition. We also removed duplicate observation and value concepts from the same healthcare organization regarding the same par-
Table 1. Classifications of family history based on the positive or negative family history, a specified disease, and a specified relative

<table>
<thead>
<tr>
<th>Class</th>
<th>Family History Classification</th>
<th>Survey Data</th>
<th>EHR Observation Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>1 Positive family history with a specified disease in specified relative</td>
<td>Responses to subquestions about relative conditions</td>
<td>Observation with relative + disease pair in a single row to confirm a match</td>
</tr>
<tr>
<td></td>
<td>2 Positive family history with a specified disease in unspecified relative</td>
<td>NA</td>
<td>EHR observations with family history disease concepts</td>
</tr>
<tr>
<td></td>
<td>3 Positive family history with an unspecified disease in specified relative</td>
<td>Responses of “Other,” “Other cancer,” or “Other/unknown diabetes” to the subquestions about relative conditions</td>
<td>Observations with family history relative concepts</td>
</tr>
<tr>
<td></td>
<td>4 Positive family history with an unspecified disease in unspecified relative</td>
<td>NA</td>
<td>Observations with “Family history of disorder” or “Family history of clinical finding” concepts</td>
</tr>
<tr>
<td>Negative</td>
<td>5 Negative family history with a specified disease in specified relative</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>6 Negative family history with a specified disease in unspecified relative</td>
<td>NA</td>
<td>Observations with “No family history of disease” concepts</td>
</tr>
<tr>
<td></td>
<td>7 Negative family history with an unspecified disease in specified relative</td>
<td>Responses of “None of the above” at the relative condition parent-question level</td>
<td>Observations with “Relative alive and well” concepts</td>
</tr>
<tr>
<td></td>
<td>8 Negative family history with an unspecified disease in unspecified relative</td>
<td>NA</td>
<td>Observations with “No family history of concepts”</td>
</tr>
<tr>
<td>Other</td>
<td>9 Other</td>
<td>Responses of “Skip”/? Prefer not to answer”/? “Don’t know”/? “Not related by blood” at the relative condition parent-question level</td>
<td>EHR observations with “Family history with explicit context”</td>
</tr>
</tbody>
</table>

EHR: electronic health record; NA: not applicable; UBR: underrepresented in biomedical research.

Statistical analysis

To compare family health history from surveys and EHRs, each survey question response or OMOP concept code was considered an observation. Each observation describes a positive or negative family history and may or may not have an associated relative or disease. A complete observation will ideally have a known family history of a specified disease for a certain relative. However, owing to the varying methods used by healthcare organizations when submitting OMOP mapped EHR data to All of Us, some information required for a complete family history may be missing. The structured nature of the family health history survey yields mostly complete observations, although the survey does not ask participants about negative family history of a specific disease. Therefore, all observations were assigned a category depending on the type of information present (Table 1). For example, there may be a positive family history or negative family history. The disease may be specified or not, meaning that the positive or negative family history is about a specific disease, such as breast cancer. Finally, a relative may be specified, such as family history in a mother. Either the disease or relative may not be specified leading to the categories in Table 1. To evaluate data types and information in the All of Us Research Program, we used descriptive statistics to (1) describe what kind of and how much information was present in the survey and EHR data, as categorized in Table 1; and (2) explore the presence of a family history of 11 diseases in the list adapted from American College of Medical Genetics and Genomics published recommendations. The descriptive statistics included frequencies and percentages of categorical data (eg, number of observations with any family history data in surveys, EHRs, or both, and observations of family history of specific diseases, such as certain cancers, heart failure, myocardial infarction, or liver disease). Two authors (A.E.H., R.M.C.) reviewed data from 10 participants to explore the overlap between the survey and EHR data.

RESULTS

General findings

There were 224 143 participants in the registered tier version of the All of Us database. Of the total participants, any EHR observation data were available in 126 252 participants. Family history–related
specific observations were available for 20,135 participants from 21 different healthcare organizations. The number of unique family history EHR observations per participant ranged from 1 to 16 (median of 1). Among the 224,143 All of Us participants who answered at least 1 question on any survey, 40,492 participants from more than 41 states completed the family health history survey (Figure 1). The number of conditions reported in survey responses ranged from 1 to 172 (median of 13). Using both data sources, there were 54,872 participants with 695,127 unique observations of any family history data. Of the total participants, 26% had family history observations from EHR only, 63% had survey observations only, and 11% had both survey and EHR observations (Figure 2). The percentage of family history data available stratified by participant-level variables including those classified as underrepresented in biomedical research, defined elsewhere (Supplementary Appendix 5), is represented in Table 2. The overall proportions of observations included in each family history category from Table 1 are summarized in Figure 3.

Survey classifications
There were a total of 658,034 family health history survey data observations (this is the denominator for the percentages reported in this section). Positive family history of a specific disease made up 87.1% observations, 3.8% identified negative family history, and 9% comprised other responses.

Positive family history
Most observations had a positive history (presence) of a disease with a relative named (Table 1, class 1: 84.9%). Owing to the structured nature of the All of Us survey data, this is the most common classification (e.g., a daughter with type 1 diabetes). A small number had only a relative named without a disease (class 3: 2.2%). Observations from this class were from survey questions that required free text responses (e.g., daughter with other cancer, please specify). Owing to the complexity of mapping free text fields to OMOP as well as the All of Us privacy rules for the public dataset that remove all free text fields, we are unable to extract the presence of a specific disease from these responses.

Negative family history
A small percentage of observations had a negative (absent) family history of a disease with a named a relative (Table 1, class 7: 3.8%).
In the survey, there is an option to select “none of the above” for all diseases in a specific relative (Supplementary Appendix 6). This is similar to responding that relative is alive and healthy.

Other
The “Other” category constituted answers submitted by participants that did not give a positive or negative family history. This 9% of observations which included the response options of “no siblings or children related by blood,” “don’t know,” or “prefer not to answer” (Supplementary Appendices 6 and 7).

EHR classifications
A total of 35,997 family history observations were from EHR data (this is the denominator for the percentages reported in this section). Positive family history made up 95.8% of the observations, 4.2% had negative family history data, and 20 or fewer observations were classified as “other.”

Positive family history
Only 1.3% (Table 1, class 1) of the family history EHR observations had both a relative and disease named. This was concluded via a row-level link in the OMOP observation table between a relative-specific family history code as an observation concept and a family history of disease code as a value concept. Most codes (59.3%) came from classes 2 and 3, indicating only a disease code or relative code without linkage. For example, sites will send a single observation disease and a separate observation for relative specification, without linking participant’s rows (Supplementary Appendix 8). In this example, a similar family history entry appeared on 2 separate dates (February 3, 2018, and February 20, 2018), and each date includes 2 codes. Of the 2 codes, one describes a family history of a disease (coronary arteriosclerosis) and the other identifies a relative (father). However, the disease and relative are not linked, so we cannot confirm that this combination refers to a father with coronary arteriosclerosis. There likely is a valid link in this case because there is only 1 disease and 1 relative present; however, linking multiple diseases and multiple relatives remains difficult. A smaller number of observations (class 4: 35.2%) identified the presence of family history, but without a disease or a relative. These concepts were “Family history of disorder” or “Family history of clinical finding.” While these data suggest that there is a family history of a condition, the specifics are unclear.

Negative family history
A small number of observations had a negative family history of a specified disease without a relative (Table 1, class 6: 1.4%). These cases included the concept code “no family history” of a disease (Supplementary Appendix 9). Very few observations had a negative family history of a relative without a disease (class 7: 1.2%).

### Table 2. Percentage of family health history information for participant-level variables including UBR as defined by the All Of Us Research Program

<table>
<thead>
<tr>
<th>Participant-Level Variable</th>
<th>Entire All of Us Cohort (n = 224,143) (%)</th>
<th>Total EHR or Survey Participants (n = 54,872) (%)</th>
<th>EHR-Only Participants (n = 14,380) (%)</th>
<th>Survey-Only Participants (n = 34,737) (%)</th>
<th>Overlap Participants (n = 5,755) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>41</td>
<td>53</td>
<td>45</td>
<td>55</td>
<td>61</td>
</tr>
<tr>
<td>Never married</td>
<td>26</td>
<td>20</td>
<td>20</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>Divorced</td>
<td>14</td>
<td>13</td>
<td>17</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Living with partner</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Widowed</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Separated</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No answer</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Health insurance status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>91</td>
<td>97</td>
<td>95</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>No</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Don’t know/No answer</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Employment status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed/self-employed</td>
<td>46</td>
<td>51</td>
<td>42</td>
<td>55</td>
<td>51</td>
</tr>
<tr>
<td>Not employed</td>
<td>51</td>
<td>48</td>
<td>56</td>
<td>44</td>
<td>48</td>
</tr>
<tr>
<td>No answer</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>UBR category (UBR definition)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall (met at least 1 UBR criterion)</td>
<td>77</td>
<td>66</td>
<td>78</td>
<td>61</td>
<td>16</td>
</tr>
<tr>
<td>Race/ethnicity (responses other than White or Hispanic/Latino)</td>
<td>48</td>
<td>26</td>
<td>42</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>Age ≥65 y</td>
<td>26</td>
<td>35</td>
<td>37</td>
<td>33</td>
<td>43</td>
</tr>
<tr>
<td>Sexual/gender minority</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Sex at birth (not male or female)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gender identity (neither man nor woman, or different than sex at birth)</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sexual orientation (responses other than straight)</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Income &lt;$25,000</td>
<td>28</td>
<td>16</td>
<td>28</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Education (&lt; GED)</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

EHR: electronic health record; UBR: underrepresented in biomedical research.
concepts described family members that were alive and well (e.g., “FH: Mother alive and well”). Finally, a negative history without a disease or relative was rarely observed (class 8: 1.7%). These were observations with a “No family history of” concept.

Other
Only 20 or fewer observations belonged to the other category, defined as having a concept of “Family history with explicit context” or “Other specified conditions influencing health status.” These codes are unclear if they signify a family history of a disorder.

Overlap between survey and EHR data
Of the 54,872 All of Us participants, 10.5% had overlapping family health history information from surveys and EHRs. To evaluate this overlap, we (1) conducted a closer examination of 10 random participants with overlapping family history observations (Supplementary Appendix 10), and (2) reviewed overlap of actionable medical disorders as presented in the next section. In all 10 cases, EHR and survey data were inconsistent, and each source provided more family health information than the other depending on the participant. Details about each participant are included in Supplementary Appendices 10 and 11.

Actionable medical disorders
There were 35,217 participants (86,054 observations) with a family history of diseases from the list of actionable medical disorders (Table 3). A family history of heart attack/myocardial infarction was the most reported (18,859 participants), followed by breast cancer (12,733 participants). All other diseases were reported by fewer than 10,000 participants. Breast cancer had the most overlap of survey and EHR data (6.2%), with all other diseases having 2% or less overlap (Table 3). EHRs contributed the most data in breast cancer (29.8% EHR only), followed by heart attack/myocardial infarction (8.7%), kidney cancer (5.7%), and colorectal cancer (3.2%). All other diseases had <1% of the information coming from EHR alone. Except for breast cancer, surveys contributed about 90% or more of the data, with congestive heart failure being almost exclusively seen in survey data.

DISCUSSION
This Demonstration Project describes and compares family health history data in surveys and EHRs mapped to OMOP from the All of Us Research Program. We discovered a large population of participants with data sourced from surveys or EHRs. There have been other registries with large populations of family health history, but they tend to focus on a specific disease or set of diseases and are usually from a single data source (e.g., surveys or EHRs).16–18 Large genetic databases like the U.K. Biobank19 and Million Veterans Program20 collect multiple data sources including genetic information. A recent study from the UK Biobank demonstrated that self-reported family history can be limited without genetics.21 However, this UK Biobank study did not report on EHR data. Our demonstration project in All of Us works with a significant amount of data from both surveys and EHRs and provides an examination of diseases with effective interventions. However, like other studies,1–7,9,10 we discovered substantial gaps in data collected from the
Table 3. Medically actionable genetic disorders from a list of genes and disorders for reporting secondary findings from genome sequencing.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Total EHR or Survey Participants (n = 54 872)</th>
<th>EHR-Only Participants (n = 14 380)</th>
<th>Survey-Only Participants (n = 34 737)</th>
<th>Overlap Participants (n = 5755)</th>
<th>Relative Ratio of Disease Discovery (EHR vs Survey)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>12 733</td>
<td>3795 (29.8)</td>
<td>8144 (64.0)</td>
<td>794 (6.2)</td>
<td>1.033</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>6753</td>
<td>28 (0.4)</td>
<td>6705 (99.3)</td>
<td>20 (0.3)</td>
<td>0.014</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>6171</td>
<td>198 (3.2)</td>
<td>5893 (95.5)</td>
<td>80 (1.3)</td>
<td>0.094</td>
</tr>
<tr>
<td>Stomach cancer</td>
<td>2033</td>
<td>20 or fewer*</td>
<td>1993 (98.0)</td>
<td>20 or fewer*</td>
<td>0.04</td>
</tr>
<tr>
<td>Thyroid cancer</td>
<td>1160</td>
<td>20 or fewer*</td>
<td>1120 (96.6)</td>
<td>20 or fewer*</td>
<td>0.071</td>
</tr>
<tr>
<td>Kidney cancer</td>
<td>1191</td>
<td>68 (5.7)</td>
<td>1103 (92.6)</td>
<td>20 or fewer*</td>
<td>0.158</td>
</tr>
<tr>
<td>Brain cancer</td>
<td>1955</td>
<td>20 or fewer*</td>
<td>1915 (98.0)</td>
<td>20 or fewer*</td>
<td>0.042</td>
</tr>
<tr>
<td>Liver disease</td>
<td>2436</td>
<td>20 or fewer*</td>
<td>2396 (98.3)</td>
<td>20 or fewer*</td>
<td>0.033</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>9595</td>
<td>59 (0.6)</td>
<td>9480 (99.2)</td>
<td>20 or fewer*</td>
<td>0.017</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>9012</td>
<td>20 or fewer*</td>
<td>8972 (100)</td>
<td>20 or fewer*</td>
<td>0.009</td>
</tr>
<tr>
<td>Heart attack/myocardial infarction</td>
<td>18 859</td>
<td>1649 (8.7)</td>
<td>16 841 (89.3)</td>
<td>369 (2.0)</td>
<td>0.236</td>
</tr>
</tbody>
</table>

Values are n (%). The percentages for each row are calculated by the total participants in each row (eg, 3795 is 29.8% of 12 733). Relative ratio of disease discovery controls the unbalanced number of participants in the EHR and survey, and is defined as:

\[
P(EHR \& Disease + | EHR) - P(Survey \& Disease + | Survey) \\
Pparticipants in EHR \times \frac{participants in the survey and Disease - }{participants in the survey and Disease - } \\
participants in EHR \times \frac{participants in the survey and Disease - }{participants in the survey}
\]

Which is equivalent to:

\[
\frac{P(EHR \& Disease + | EHR) - P(Survey \& Disease + | Survey)}{Pparticipants in EHR \times \frac{participants in the survey and Disease - }{participants in the survey}}
\]

For breast cancer, this is \(\frac{1.033}{0.014} = 1.033\). A relative risk > 1 means that the EHR contains more disease information; a relative risk < 1 means that the survey contains more disease information.

EHR: electronic health record; UBR: underrepresented in biomedical research.

\(a\)According to All of Us publication policy, all counts of participants in publications that are less than 20 should be generalized to the phrase “20 or fewer.”

EHR and surveys. There are many important reasons for these gaps. Mapping of EHR data to a structured common data model (eg, OMOP) can lead to a loss of information because of an inability to obtain unstructured data from the EHR or difficulty transforming structured data to the common data model,\(^{22,23}\) such as error or omission of family history data due to participants forgetting information or not being aware of the information, or healthcare providers not obtaining or documenting the information. Finally, a lack of an ability to enter data, such as limiting the number of questions on a survey or what structured data can be entered on a specific EHR system, can cause these gaps. There were also interesting findings in the overlap of survey and EHR data. Despite limitations, our findings suggest that multiple data sources may provide more accurate family health history information than a single source alone.

Interesting insights surfaced from the examination of overlapping survey and EHR data. On the one hand, in many cases, more family history was available in the survey responses as compared with the EHR. The structured nature of the survey offers an opportunity for a more systematic collection of family history of specific diseases on relatives. On the other hand, diseases in EHRs were sometimes missing from the survey records. This discrepancy ultimately creates uncertainty in determining the gold standard for family health history information. In 1 case, a participant had responded that their father had a heart attack in the family health history survey (Supplementary Appendix 12), but their EHR noted that their father had coronary artery disease. Reasons for discrepancies between EHR and survey data could include (1) a chronological gap between when the participant answered the survey question and when the EHR data were entered (eg, a father may have had coronary artery disease at the time of the survey and later had a heart attack), (2) the ability of the participant to remember or understand the father’s disease, or (3) a data entry error by either the healthcare provider or the participant. Resolving these sources of discrepant information are areas for future research. Additional data sources can provide important complementary information. All of Us is collecting DNA for future genetic analyses. A comprehensive examination of the genome for pathogenic or likely pathogenic gene variants or polygenic risk scores in participants who report a family history may complement data from the patient surveys and EHR. Additionally, answers to questions such as “what percent of individuals reporting or not reporting family history of breast cancer have a mutation promoting breast cancer?” could be explored.

Medically actionable diseases also had discrepancies between the EHRs and surveys. History of breast cancer was present in more relatives in the EHR data than surveys because the family health history survey does not ask about distant relatives. Information about a participant’s family history of gastric cancer was present in the survey but not in the EHR, possibly owing to issues with specificity of what is in the EHR data or mapping of raw data to the OMOP format (eg, neoplasm of digestive organ). Therefore, ensuring that data in the EHR have the appropriate specificity of the disorder (eg, gastric cancer and not neoplasm of digestive organism) as well as improving mapping to OMOP can help. There was almost no overlap between surveys and EHR in the liver disease category (fewer than...
The survey asks about a family history of liver disease, but again EHR data may not have been mapped to an OMOP concept like cirrhosis or liver disease because it is located in unstructured data or not typically captured. Generally, survey data had more extensive family history. In 1 participant’s family history of thyroid cancer, the EHR had only 1 mention of thyroid cancer for a specified relative, whereas the survey had 3 relatives with thyroid cancer. Owing to the lack of disease-relative linkage guidelines in the mapping of EHR data to OMOP, the surveys had more specific family history information. When considering certain mutations such as BRCA in analyzing the family history of breast cancer, more specific family history will be important. The finding that most medically actionable genetic disorders were in the surveys but not EHRs emphasizes the importance of properly capturing and mapping family history to common data models.

Data quality of family history within All of Us has important limitations and considerations. First, we do not have access to free text from the EHR or free text answers in the survey due to privacy rules in the registered tier to prevent reidentification of participants. Future directions to utilize natural language processing and map to OMOP could enable these data to be used. Second, limitations within EHR-specific family health history included a lack of linkage between diseases, relatives, and duplicates. Only a few sites sent linked OMOP codes between diseases and relatives. This could be related to the way the information is stored in their EHR or the lack of awareness that diseases and relatives could be transmitted to the data repository as a linked set. Improving the standard way in which these data are collected, transmitted, and stored for All of Us researchers could allow for more extensive analyses. Also, using a different standard like the Fast Healthcare Interoperability Resources could lead to different results. Third, the response rate of the family health history survey is low, and the All of Us Research Program is engaged in a campaign to increase completion rates of surveys from nonresponders through phone calls and postal mailings. Response rates in surveys were particularly low in certain underrepresented in biomedical research populations such as lower education. Statistical weighting methods, such as inverse probability weighting, multiple imputation, or a combination of the 2, could be useful when trying to adjust for nonresponders. In addition, there may be other confounders limiting survey data, such as digital literacy, availability of stable Internet connections, or smartphones. These limitations should be considered by researchers using these data. Fourth, we assume that responses from surveys and EHR family history data are from blood relatives; however, participants may respond about adoptive relatives. Fifth, there were a significant number of duplicates in the EHR data. These duplicates may be true duplicates or could be additional family history. This distinction was difficult to determine given the limited row-level linkages between EHR diseases and relatives. True duplicates are potentially attributable to data being recorded in the EHR systems used to populate the OMOP instance at a 1-row-per-contact basis. Therefore, every time the participant completes a family history form or every time a provider reviews the family history data, another row could be populated in the EHR source data, and thus mapped into OMOP and sent to the data repository. Sixth, including a survey with all possible diseases and relatives may give more data but increases participant burden. Survey responses are also susceptible to misclassification and multiple biases. Last, these results are specific to the release of the registered tier in September 2019; future releases may include additional family history data (eg, additional OMOP codes) and data types not included in the current analysis.

The Demonstration Projects led by All of Us aims to describe the utility of family health history data available to researchers at the beta launch of the Researcher Workbench. All analyses shown here are available to registered users of the Researcher Workbench for replication and reuse to support hypothesis generation and discovery by the broader research community.

**CONCLUSION**

This description of the family health history data in the All of Us registered tier database will assist future investigators in understanding All of Us data methods and provide feedback to the program on the utility of participant surveys and EHR data. In this Demonstration Project, we demonstrated the potential informatics challenges and opportunities for biomedical research of family history data from different sources, which were mapped to a common data model in an attempt to identify a common source of truth regarding family history in a large, diverse cohort of participants across the United States.

**FUNDING**

The All of Us Research Program is supported (or funded) by grants through the National Institutes of Health Office of the Director: Regional Medical Centers (1 OT2 OD026549, 1 OT2 OD026554, 1 OT2 OD026557, 1 OT2 OD026556, 1 OT2 OD026550, 1 OT2 OD026552; 1 OT2 OD026553, 1 OT2 OD026548, 1 OT2 OD026551, 1 OT2 OD026555, IAA #: AOD 1603); Federally Qualified Health Centers (HHSN 263201600085U), Data and Research Center (5 U2C OD023196), Biobank (1 U24 OD023121), the Participant Center (U24 OD023176), Participant Technology Systems Center (1 U24 OD023163), Communications and Engagement (3 OT2 OD023205, 3 OT2 OD023206), and Community Partners (1 OT2 OD023277, 3 OT2 OD023315, 1 OT2 OD023337, 1 OT2 OD025276). In addition to the funded partners, the All of Us Research Program would not be possible without the contributions made by its participants. RMC was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number K23HL141447. JHK is supported by the National Institutes of Health Office of the Director (1OT2OD026549) and the National Heart, Lung, And Blood Institute of the NIH (K01HL143137).

**AUTHOR CONTRIBUTIONS**

RMC, RJC, CJO, AHR conceived the study design. All authors contributed to data collection. RMC, AEH, XF, QC performed data analysis and interpretation of the results. All authors contributed to the writing and review of the manuscript.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at *Journal of the American Medical Informatics Association* online.
ETHICAL APPROVAL

The Institutional Review Boards of the All of Us Research Program approved all study procedures and informed consent was obtained from all participants.

DATA SHARING

De-identified data are available on the researcher workbench of the All of Us Research Program located at https://workbench.researchalofus.org.

ACKNOWLEDGMENTS

The authors want to thank the participants of the All of Us Research Program. We thank the countless other co-investigators and staff across all awardees and partners, without which All of Us would not have achieved our current goals. “Precision Medicine Initiative, PMI, All of Us, the All of Us logo, and The Future of Health Begins with You are service marks of the U.S. Department of Health and Human Services.”

The All of Us Research Program Demonstration Projects Subcommittee includes the following authors: Cheryl Clark; Elizabeth Cohn; Lucila Ohno-Machado; Mine Cicek; Eric Boerwinkle; Sheri Cohn; Steve Mockrin; Kelly Gebo.

REFERENCES