Oregon All Payer All Claims (APAC) Program

Application for Limited Data Files

APAC-3

This application is used to request limited data sets. If you would like to discuss APAC data in relation to your project prior to submitting this application, please contact apac.admin@state.or.us with a brief description of the project and your contact information. OHA will have someone contact you to help determine if APAC is appropriate for your project and, if so, which data elements may be needed.

PROJECT INFORMATION

Project Title: Health equity and the impacts of EHR data bias associated with social determinants

Principal Investigator: Nicole G. Weiskopf

Title of Principal Investigator: Associate Professor

Organization: Oregon Health & Science University

Address: 3181 SW Sam Jackson Park Rd

City: Portland State: OR Zip Code: 97239

Telephone: 347-216-0520

Email: weiskopf@ohsu.edu

SECTION 1: PROJECT SUMMARY

1.1 Project Purpose: Briefly describe the purpose of the project. You may submit a separate document that details the project's background, methodology and analytic plan in support of your request for APAC data elements.

Please see additional response document.

1.2 Research Questions: What are the project's key research questions or hypotheses? If this project is research and has been approved by an Institutional Review Board (IRB), the research questions must align with the IRB approval documentation. If needed, a more detailed response may be submitted as a separate file.

- Note: APAC staff will use your response to this question to determine the minimum data elements necessary for this project, in accordance with the HIPAA minimum necessary standard. The research questions should be specific enough to justify the need for each data element beyond identifying it as a "potential confounding variable."

Please see additional response document.

1.3 Products or Reports: Describe the intended product or report that will be derived from the requested data and how this product will be used. If needed, a more detailed response may be submitted as a separate document with this application.

Please see additional response document

1.4 Project Timeline: What is the timeline for the project?

Anticipated Start Date: 08/04/2023

Anticipated Publication/Product Release Date: TBD

Anticipated End Date: 04/30/2027

1.5 Data files may not be released or reused beyond the terms of the data use agreement resulting from this application regardless of funding source or other obligations of the principal investigator, organization or research team.

☒ I understand this limitation and agree that data files or work products will not be shared at less than an aggregated, de-identified level.

☐ I understand this limitation and request approval to share data files or work products at a potentially re-identifiable level as follows:

SECTION 2: PROJECT STAFF

2.1 Project Staff: Please list all individuals in addition to the principal investigator who will have direct or indirect access to the data. This must include any contractors or other third parties with access to the data.

Name: Nicole Weiskopf Email: weiskopf@ohsu.edu	Project role: OHSU Principal Investigator
Name: Teresa Schmidt Email: schmidtt@ochin.org	Project role: OCHIN Site-PI
Name: Caroline Thompson Email: caroline.thompson@unc.edu	Project role: Co-Investigator
Name: Steven Bedrick Email: bedricks@ohsu.edu	Project role: Co-Investigator
Name: Ana Quiñones Email: quinones@ohsu.edu	Project role: Co-Investigator
Name: David Dorr Email: dorrd@ohsu.edu	Project role: Co-Investigator
Name: Matthew Jones Email: jonestm@ochin.org	Project role: Research Analyst

Attach additional sheets as needed.

2.2 Technical Staff: Please list any additional staff who will be maintaining the data file(s) or otherwise assisting in the transfer or receipt of the data files. Files will not be transferred to anyone who is not listed on this application as either project staff or technical staff.

Name: Email:	Technical role:
Name: Email:	Technical role:

Attach additional sheets as needed.

SECTION 3: DATA REQUEST

3.1 Purpose of the Data Request:

a. Listed below are the purposes for which OHA may share APAC data. Please choose the category in which your project falls under (***choose only one***).

- [■] Research (refer to 45 CFR 164.501 for definition)
- [ ] Public health activities as defined in 45 CFR 164.512(b) by the state or local public health authority
- [ ] Health care operations as defined in 45 CFR 164.501
  Covered entity as defined in 45 CFR 160.103? [ ] Yes [ ] No
- [ ] Treatment of patient by health care provider as defined in 45 CFR 164.506 (c)(2)
  Covered entity? [ ] Yes [ ] No
- [ ] Payment activities performed by covered entity or health care provider as defined in 45 CFR 164.506 (c)(3)
  Covered entity? [ ] Yes [ ] No
- [ ] Work done on OHA's behalf by a Business Associate as defined in 45 CFR 160.103

b. Describe how the project falls into the category chosen above.
Please see additional response document.

3.2 Direct identifiers. What level of data identifiers are you requesting (***choose only one***)? Reference the Data Elements Workbook for the categorization of data elements.

- [■] De-identified (as outlined in 45 CFR 164.514(e)) protected health information
- [ ] Limited, potentially re-identifiable data elements
- [ ] Restricted direct identifiers (member name, address, date of birth, etc.) *Please note:* Direct identifiers are only released under special circumstances that comply with HIPAA requirements, and will require specific approvals, such as IRB approval, patient consent and/or review by the Oregon Department of Justice.

3.3 Human Subjects Research: IRB protocol and approval are required for most research requests for limited data elements. Not obtaining IRB approval or waiver in advance may delay approval of the data request. **The research questions reported in 1.2 of this application must match the documentation supporting the IRB approval received or the IRB approval will not be accepted for this data application.**

The IRB application should indicate that APAC data contains sensitive personal health information and is subject to HIPAA regulations.

a. Does the project have IRB approval for human subjects research or a finding that approval is not required?

☒ Yes ☐ No

If no, briefly explain why you believe that this project does not require IRB review.

If an IRB reviewed the project, include the IRB application and approval/finding memo with the submission of this APAC-3 and complete parts b-e below.

☑ IRB application and approval memo are attached.

b. Describe how this application is within the authority of the approving IRB.

Please see additional response document.

c. Describe why the project could not be practicably conducted without a waiver of individual authorization (a waiver of individual authorization is provided by the IRB in cases in which the researcher does not need written authorization from participants to use their PHI):

Please see additional response document.

d. On what date does the IRB approval expire? 7/10/2026

SECTION 4: DATA ELEMENTS

4.1 Narrowing Data Needs: Refer to the APAC Data Dictionary for detailed information about the data elements. In compliance with HIPAA regulations, you will only receive data elements that are adequately justified. This means APAC will only provide the minimum necessary data required for the project as represented in the research questions, protocol and IRB approval.

a. What years of data are requested? 2011 through 2021 are currently available.

2011-2022

b. What payer types are requested? Check all that apply

☑ Commercial ☑ Medicaid ☑ Medicare Advantage

c. What types of medical claims are requested? ☑ All

☐ Inpatient hospital ☐ Emergency department ☐ Outpatient
☐ Ambulatory surgery ☐ Ambulance ☐ Transportation
☐ Hospice ☐ Skilled Nursing Facility ☐ Professional

d. Demographic data limitations

1. Gender ■ All ☐ Male ☐ Female

2. Age ☐ All ☐ Only 65+ ☐ Only 18 and younger ■ Other 18+ by end of 2022
(Specify age range)

e. Will data requested be limited by diagnoses, procedures or type of pharmaceutical? Add additional sheet if needed.

Diagnoses, indicate ICD 9 and ICD10 codes to include:
A list will of ICD codes will be provided separately

Procedures, indicate CPT to include:
All (for study patients meeting global criteria)

Pharmaceuticals, indicate NDC or therapeutic classes to include:
A list of NDC codes will be provided separately

f. APAC has a small number of out-of-state residents included, most often through PEBB or OEBB coverage. Do you want to include out-of-state residents? ☐ Yes ■ No

4.2 Data Element Workbook: Complete the Data Element Workbook to identify specific data requested.

Data Element Workbook completed and attached, including justifications for each element requested. ☑

SECTION 5: DATA MANAGEMENT & SECURITY

5.1 Data Reporting: APAC data or findings may not be disclosed in a way that can be used to re-identify an individual. Data with small numbers – defined as values of 30 or less (n≤30) or subpopulations of 50 or fewer individuals (n≤50) – cannot be displayed in findings or outputs derived from APAC data. Please describe the techniques you will use to prevent re-identification when findings or outputs result in small numbers or subgroups (e.g. aggregation, cell suppression, generalization, or perturbation).

OCHIN has a company-wide requirement to suppress patient counts between 1 and 10 (inclusive) in all research dissemination activities. We employ automated and manual checks by project teams before each dissemination or disclosure. When using APAC data, we will adjust this method to suppression counts 1-30 for the larger population and 1-50 for sub-populations for any dissemination or disclosure.

5.2 Data Linkage: OHA seeks to ensure that APAC data cannot be re-identified if it is linked or combined with data from other sources at the record, individual or address level. Requesters are strongly encouraged to consult with APAC staff regarding linking APAC data with other data prior to submitting a data request. Health Analytics prefers to conduct APAC data linking in-house and share only encrypted identifiers with data requesters.

a. Does this project require linking to another data source?

■ Yes ☐ No

If yes, please complete parts b-d below.

b. At what level will data be linked?

☐ Address ☐ Facility ✔ Individual person/member

☐ Individual provider

c. If required to link

■ Authorized to provide data for linking at OHA

☐ Not authorized to provide data for linking at OHA

☐ Unknown

d. Describe and justify all necessary linkages, including the key fields in each data set, how they will be linked, the software proposed to perform the linkage and why it is necessary.

Please see additional response document.

e. Describe in detail the steps will you take to prevent re-identification of linked data.

OCHIN will send a 'finder file' of personal identifiers to HSRI/NORC for linkage to APAC persons. This finder file will contain requested PII fields and also an OCHIN Client_ID that is a randomized string 36 characters in length with no PHI or PII. OCHIN requests a crosswalk from OCHIN Client_ID to APAC Member_ID be provided to OHSU, along with other APAC datasets. This crosswalk will be instructed not to contain any PII fields to prevent re-identification of linked data. OCHIN will otherwise be sending OHSU limited datasets (with no direct identifiers) regarding health services received in OCHIN clinics. OHSU will link OCHIN datasets to the APAC datasets through use of the crosswalk.

5.3 Data Security (required for all applications):

a. Attach a detailed description of your plans to manage security of the APAC data including:
   - Designation of a single individual as the custodian of APAC data, either the principal investigator or staff listed in Section 2 of this application, who is responsible for oversight of APAC data, including reporting any breaches to OHA and ensuring the data are properly destroyed upon project completion.
   - A security risk management plan applicable to APAC data that includes:
     - Secure storage in any and all mediums (e.g., electronic or hard copy)
     - Procedures to restrict APAC data access to only those individuals listed on the data use agreement
     - User account controls, i.e., password protections, maximum failed login attempts, lockout periods after idle time, user audit logs, etc.
     - Confirmation of training for personnel on how to properly manage protected health information in all formats
     - Protection of derivatives of APAC data at the re-identifiable level
     - If applicable, procedures for handling direct identifiers, such as allowing access on a 'need to know' basis only and minimizing risk by storing identifiers separately from other APAC data
     - Procedures for identifying, reporting and remedying any data breach
   - Statement of compliance with HIPAA and the HITECH Act
   - Electronic device protections, i.e., anti-virus or anti-malware software, firewalls, and network encryption

b. Record level or derivative data that can be re-identified must be destroyed within 30 days of the end of the data use agreement, in a manner that renders it unusable, unreadable or indecipherable. What are your plans for destruction of the dataset and any potentially identifiable elements of the data once the data use agreement has expired?

Please see additional response document.

SECTION 6: COST OF DATA

Because each data set is unique, cost can be determined only after the specific data elements are finalized. APAC staff will then review your request and estimate the number of hours required to produce and validate the data. APAC is currently requiring reimbursement for the cost of file transfer only ($890 per request). Payment must be received before the data will be provided. APAC staff will provide an invoice to facilitate payment. OHA's W-9 is available on request.

SECTION 7: CHECKLIST AND SIGNATURE

7.1 Checklist: Please indicate that the following are completed:

- ■ I acknowledge that payment will not be refunded if OHA fulfills the data request, but the receiving entity does not have the capability to import or analyze the data
- ■ All questions are answered completely
- ■ Data Element Workbook is attached to email or printed application
- ■ IRB application with approval/finding memo is attached to email or printed application, if applicable
- ■ Data privacy and security policies for the requesting organization, and any third-party organizations, are attached to the email or printed application

7.2 Optional Racial Justice Addendum: Please see the last two pages of this form for options if data will be used to eliminate racial injustice.

- ■ I am interested in this option
- ☐ This option does not apply to my data request

7.3 Signature: The individual signing below has the authority to complete this application and sign on behalf of the organization identified in Section 1. By signing below, the individual attests that all information contained within this data Request Application is true and correct.

Signature Date 03/20/2024

Printed name Nicole Weiskopf

Title Associate Professor, Medical Informatics and Clinical Epidemiology

Return the completed form with required attachments to APAC.Admin@odhsoha.oregon.gov.

Optional APAC Addendum: Using APAC Data to Eliminate Racial Injustice

Requestors may complete this optional section if their project will identify concrete actions to eliminate health inequities stemming from historical and contemporary injustices and the inequitable distribution of resources and power (see Health Equity definition on next page). For projects that inform such solutions, and **do not simply document disparities**, the Director of the **Office of Health Analytics** may, at their discretion, offer one or more of the following incentives:

- Priority processing of requestor's application
- Waiver of fees
- Priority production of data files
- Technical assistance from APAC analysts
- Access to enhanced race and ethnicity data in the future. (Race/ethnicity data in APAC are currently limited because entities that submit administrative data to APAC do not generally include race/ethnicity information.)
- Other provisions that the Director of Health Analytics may find appropriate

Receipt of any of these incentives requires requesters to deliver to the Office of Health Analytics a document fully describing the analytic methods at the conclusion of the relevant analyses, including:

- Commercial off-the-shelf applications used
- Grouping and aggregation methods
- Algorithms and calculations
- Use of code sets that are proprietary to a third party not associated with the project
- Copies of programming code attached in an appendix

The Office of Health Analytics will compile a compendium of analytic methods and make this freely available on the APAC web site. Requestors are also encouraged to submit copies of publications or products using the APAC data for posting on the APAC web site. See below for additional information and application instructions.

Using APAC Data to Eliminate Health Inequities

Problem: Health inequities due to institutional racism and racial injustice

Solution: Develop methods for using APAC data to eliminate institutional racism and racial injustice.

Goal: Eliminate institutional racism and racial injustice, including discrimination based on the intersections of race, ethnicity, language and disability.

Rationale: OHA recognizes that historical and contemporary racial injustice is a root cause of health inequity. APAC and its users, who have subject matter expertise, infrastructure, and staffing sufficient to use the large and complex data files, comprise a community of privilege. As such, APAC has an obligation to use its privilege to confront institutional racism and racial injustice, within OHA specifically and across Oregon. The APAC community has a tremendous wealth of research expertise that could develop novel methods for using APAC data to document racial injustice and identify opportunities to eliminate it.

Instructions: In a separate attachment, describe in detail:

- How requestor's research will help requestor's organization and OHA document racial injustice **and** identify opportunities to eliminate it. Requestor's description must be thorough and as specific as possible and should describe how the research findings will be consistent with OHA's efforts to achieve true Health Equity (see definition, below). **Simply documenting disparities is not sufficient**.
- How requestor's research will be explicitly clear and open about the methods used, widely replicable, and not proprietary to requestor's organization or to a third party. Note that this does not preclude requestor's use of necessary codes sets, such as CPT codes, that are proprietary to a third party and available for license.
- How requestor's organization will freely share the key findings.

A note on intersectional research into inequities based on race, ethnicity, language and disability: Researchers are encouraged to consider an intersectional approach that encompasses language and disability when researching strategies to eliminate racism and racial injustice. However, administrative claims data submitted to APAC generally do not include data on language or disability. APAC includes some race and ethnicity data, but it encompasses less than half of the people in the database. To mitigate these limitations, OHA staff may be able to provide assistance to selected applicants interested in intersectional approaches, as staff resources permit.

Health Equity Definition

Oregon will have established a health system that creates health equity when all people can reach their full health potential and well-being and are not disadvantaged by their race, ethnicity, language, disability, gender, gender identity, sexual orientation, social class, intersections among these communities or identities, or other socially determined circumstances.

Achieving health equity requires the ongoing collaboration of all regions and sectors of the state, including tribal governments to address:

- The equitable distribution or redistributing of resources and power; and
- Recognizing, reconciling and rectifying historical and contemporary injustices.

Oregon All Payer All Claims (APAC) Program
Application for Limited Data Files
APAC-3

Additional Response Document

Project Title: Social Determinants of Health and Clinical Prediction Bias

Principal Investigator: Nicole Weiskopf, PhD; Associate Professor of Medical Informatics and Clinical Epidemiology

1.1 Project Purpose: Briefly describe the purpose of the project.

We are conducting a systematic exploration of the association between social determinants of health (SDoH) and electronic health record (EHR) data quality, and we are measuring how EHR data quality affects the performance of several standard clinical risk prediction algorithms. We hypothesize that EHR data quality varies by SDoH, and that clinical risk prediction tools perform worse when EHR data quality is lower. To investigate these associations, we will use EHR and SDoH data from OCHIN – a non-profit health information and innovation network that supports the nation's largest network of outpatient safety net practices using a single EHR (see www.ochin.org). However, the OCHIN EHR is largely restricted to care delivery in outpatient settings and does not include data from hospitals or specialty clinics. So, in this APAC-3 application we propose using All Payer All Claims data from the state of Oregon to (1) identify key inpatient clinical outcomes: hospital readmissions, cardiovascular diseases and events, and (2) improve our capture of select patient conditions and medications that inform clinical risk prediction algorithms.

OHSU has previously acquired APAC data on OCHIN patients for another research project (APAC request #5743), through patient linkage between APAC and OCHIN's EHR. We would like to use that same methodology to acquire APAC data for the calendar years 2011-2022, to address the following project aims:

- Characterize relationships between SDoH, EHR data quality (i.e., completeness of capture of patients' underlying medical conditions), and healthcare utilization.
- Determine the association between risk prediction validity (defined as the accuracy of predicting occurrences of hospital admissions, cardiovascular diseases and events) and EHR data quality, as well as the association between risk prediction validity and SDoH.
- Leverage existing debiasing methods to protect the validity of clinical prediction and risk assessment using EHR data.

Our results will provide conceptual and methodological solutions to ensure that any task relying on the reuse of EHR data is equally valid for all patient populations regardless of their SDoH. While this proposal focuses on clinical risk prediction, our findings will be applicable to analytical tasks throughout the learning health system, including patient care, quality improvement, and research.

1.2 Research Questions: What are the project's key research questions or hypotheses?

Our first aim in linking APAC and OCHIN EHR data will be to characterize the relationships between social determinants of health (SDoH), healthcare utilization, and EHR data quality. EHR data quality will include examination of completeness, conformance (adherence to standards), and plausibility (accuracy) of standard intake fields as well as fields of particular relevance for six clinical algorithm use cases.

- We will rely on the OCHIN EHR to derive patients' SDoH information, including demographics, insurance type, income level, SDoH screening results, and area-level environmental factors from the American Community Survey and US Census that OCHIN has linked to patients based on geocoded patient addresses.
- We will use APAC data to measure some aspects of OCHIN's EHR data quality, by checking it for additional underlying medical conditions that should inform clinical risk prediction algorithms but were not indicated in OCHIN's EHR.
- We will use both OCHIN EHR and APAC data to evaluate healthcare utilization, to understand whether the frequency and modality of care utilization are associated with EHR data quality. If associated with data quality and SDoH, we will test whether healthcare utilization mediates the relationship between SDoH and EHR data quality.
- We will look generally at data completeness across standard patient intake fields (e.g., demographics, anthropometrics), as fields related to six clinical prediction algorithm "use cases": LACE score, HOSPITAL score, Atherosclerotic Cardiovascular Disease (ASCVD) risk calculator, Simple Framingham Risk Score, Charlson Comorbidity Index, and Elixhauser Comorbidity measure.
- Our specific hypotheses for Aim 1 include the following:
  H1a: EHR data completeness is lower for underserved and at-risk patients.
  H1b: Healthcare utilization mediates the association between overall EHR record completeness and patient SDoH.
  H2a: Use case-specific EHR data completeness is lower for underserved and at-risk patients.
  H3b: Healthcare utilization mediates the association between use case-specific EHR data quality and patient SDoH.

Our second aim in linking APAC and OCHIN EHR data will be to determine SDOH and EHR data quality are associated with risk prediction validity, using clinical algorithms that are used to predict hospital readmissions (LACE, HOSPITAL), cardiovascular diseases and events (ASCVD, Framingham), and mortality (Charlson, Elixhauser).

- We will need APAC data on inpatient stays (admission type, length of stay, inpatient procedures, oncology services, etc.) to help us measure the accuracy of the HOSPITAL score and LACE score for risk prediction of hospital readmission.
- We will need APAC data on cardiovascular diseases and events (diabetes, hypertension, statin use, aspirin therapy, etc.) to measure the accuracy of the Atherosclerotic Cardiovascular Disease (ASCVD) risk calculator and the Simple Framingham Risk Score.
- We will need APAC data on medical conditions (stroke, dementia, lymphoma, etc.) to help us measure the accuracy of the Charlson Comorbidity Index and Elixhauser Comorbidity measure for predicting mortality. OCHIN's EHR contains death data from the Social Security Death Master File and US obituary data, acquired through Datavant.
- Our specific hypotheses for Aim 2 include the following:
  H1: Calculated risk scores are significant predictors of observed clinical outcomes.
  H2a: Data are less likely to be sufficient for risk score calculation for underserved and at-risk patients.

H2b: Data are less likely to be sufficient for risk score calculation for patients with lower overall and use case-specific data quality.
H3a: For patients with sufficient data for risk score calculation, variation in predictive accuracy is partially explained by SDoH and utilization.
H3b: For patients with sufficient data for risk score calculation, variation in predictive accuracy is partially explained by overall and use case-specific data quality.

Our third aim in linking APAC and OCHIN EHR data will be to leverage existing debiasing methods to protect the validity of clinical prediction and risk assessment for patients with lower levels of data quality in the EHR. This aim will not require any additional APAC data.

- Our specific hypotheses for Aim 3 include the following:
  H1a: Causal models of risk scores incorporating data quality and SDoH concepts are significant predictors of actual outcomes.
  H1b: Causal models of risk scores incorporating data quality and SDoH concepts are better predictors of actual outcomes than the original risk scores.
  H2: Failure of risk models can be predicted using a set of discrete factors that are easily assessed at the point of care.

1.3 Products or Reports: Describe the intended product or report that will be derived from the requested data and how this product will be used.

This project will generate academic publications and dissemination activities in academic conferences and related venues. Results from the analyses conducted in aims 1 and 2 of this project will promote awareness of how social determinants of health (SDoH), healthcare utilization, and healthcare data quality impact the performance of clinical risk prediction algorithms. We hope the debiasing methods developed in aim 3 of this project will inform future improvements to risk prediction algorithms, such as flagging patients at the point of care when algorithms may be underperforming or allowing providers additional interventions and/or information gathering opportunities to protect algorithm performance in situations with suboptimal data quality.

2.1 Project Staff: Please list all individuals in addition to the principal investigator who will have direct or indirect access to the data. This must include any contractors or other third parties with access to the data.

Name: Nicole Weiskopf Email: weiskopf@ohsu.edu	Project role: Principal Investigator
Name: Teresa Schmidt Email: schmidtt@ochin.org	Project role: OCHIN Site-PI
Name: Caroline Thompson Email: caroline.thompson@unc.edu	Project Role: Co-Investigator
Name: Steven Bedrick Email: bedricks@ohsu.edu	Project Role: Co-Investigator
Name: David Dorr Email: dorrd@ohsu.edu	Project Role: Co-Investigator
Name: Ana Quiñones Email: quinones@ohsu.edu	Project Role: Co-Investigator
Name: Matthew Jones Email: jonesm@ochin.org	Project Role: Research Analyst
Name: Rae Crist Email: cristr@ochin.org	Project Role: Research Associate
Name: Lily Cook Email: cookli@ochin.org	Project Role: Research Analyst
Name: Christie Jackson Email: jackschr@ohsu.edu	Project Role: OHSU Project Manager

3.1 Purpose of the Data Request:
b. Describe how the project falls into the category chosen above.

We intend to pursue several NIH-funded study aims (R01LM013990) by combining APAC data OCHIN EHR data and linked data resources (described in section 5.2.d).

We hope our results will inform approaches to improve clinical risk prediction for all patient populations, but we are not conducting public health activities or providing any healthcare in this project. We do not plan to have any interaction with the patients involved in this research. We are conducting data-only studies in this work.

3.3 Human Subjects Research: IRB protocol and approval are required for most research requests for limited data elements. Not obtaining IRB approval or waiver in advance may delay approval of the data request.

b. Describe how this application is within the authority of the approving IRB.

The project was approved by the OHSU IRB (FWA00000161; STUDY00025986) according to NIH JIT policy and determined that it would be Expedited under Category # 5. In the study, we use secondary data only. The research will not involve any further contact with human subjects to collect data. It is based on existing clinical and administrative data that were collected, compiled, and stored prior to the initiation of the project. The OHSU IRB complies with 45 CFR Part 46, 21 CFR Parts 50 and 56, and other federal and Oregon laws and regulations, as applicable, as well as ICH-GCP codes 3.1-3.4, which outline Responsibilities, Composition, Functions, and Operations, Procedures, and Records of the IRB. The research is not classified, is not regulated by the FDA and does not involve prisoners.

c. Describe why the project could not be practicably conducted without a waiver of individual authorization (a waiver of individual authorization is provided by the IRB in cases where the researcher does not need written authorization from participants to use their PHI):

Secondary data pertains to a significant number of individuals (~100,000), so acquiring authorization from each person would not be practical. The data to be used for analysis are already collected, so their acquisition does not pose greater risk of harm encountered than that in daily life. This research involves risk of breach of confidentiality or unauthorized disclosure, but appropriate protections for subject privacy and confidentiality are in place.

5.2 Data Linkage: OHA seeks to ensure that APAC data cannot be re-identified if it is linked or combined with data from other sources at the record, individual or address level. Requesters are strongly encouraged to consult with APAC staff regarding linking APAC data with other data prior to submitting a data request. Health Analytics prefers to conduct APAC data linking in-house and share only encrypted identifiers with data requesters.

d. Describe and justify all necessary linkages, including the key fields in each data set, how they will be linked, the software proposed to perform the linkage and why it is necessary.

We will follow OHA's suggestion and request assistance from HSRI/NORC for linking OCHIN patients to APAC claims records. This linkage is critical for us to be able to study the connection between patients' social determinants of health (SDoH; as documented in our EHR data) and the accuracy of clinical prediction algorithms in predicting hospital readmissions and cardiovascular disease and events (as evident in APAC claims).

OCHIN's EHR data are already linked to dates of death from Datavant through an encrypted algorithmic process where no PIDs are shared between us and Datavant. Datavant death data include records from the Social Security Administration, as well as current and historical obituary data. OCHIN's EHR data on patient addresses are also linked, in an ongoing basis, to area-level indicators from the US Census and American Community Survey (ACS) that are publicly available. OCHIN geocodes patient addresses and maps them geographic areas so we can reference environmental attributes through those publicly available data sources. (Some previous patient addresses were geocoded and linked to Census and ACS data by HealthLandscape at the Robert Graham Center.) OCHIN linkages to these additional data resources are critical for us to investigate our aims.

We will not be linking APAC data to any other datasets aside from those that are already linked to our OCHIN EHR.

Specifically, we'd like to request APAC claims data for study patients who meet any of the following criteria, and a list of study patients who do not meet any of these criteria:

1. Please provide all claims 2011-2022 for patients with any diagnosis from a list of diagnosis codes that we will provide (for the Charlson Comorbidity Index and Elixhauser Comorbidity Measure).
2. Please provide all claims 2011-2022 for patients with any prescription from a list of relevant drugs we will provide NDC codes for (including statins, aspirin, and HTN medications).
3. Please provide claims 2019-2022 for patients with claims to indicate an inpatient admission or an ED visit, based on APAC groupers for defining inpatient visits and ED visits

We would also like to request that the patient crosswalk (from OCHIN Client_ID to APAC member_ID) include member_IDs for patients who did not meet any eligible criteria. That will help us to differentiate which of our OCHIN study patients did not link to any APAC member IDs, versus those who linked to someone but did not meet any of the above inclusion criteria.

5.3 Data Security (required for all applications):

a. Attach a detailed description of your plans to manage security of the APAC data including:

Designation of a single individual as the custodian of APAC data, either the principal investigator or staff listed in Section 2 of this application, who is responsible for oversight of APAC data, including reporting any breaches to OHA and ensuring the data are properly destroyed upon project completion.

Dr. Nicole Weiskopf, study PI, will act as the data custodian for the Oregon All Payer All Claims data. Dr. Weiskopf will oversee the management and use of the APAC data in accordance with the Data Use Agreement (DUA). Additionally, she will be responsible for reporting any breaches and will ensure that all data are properly destroyed upon completion of the proposed project and in accordance with the DUA.

A security risk management plan applicable to APAC data that includes:

1. Secure storage in any and all mediums (e.g., electronic or hard copy)
   All OHSU removeable media is encrypted. Dell Data Protection is required to connect a Windows computer to OHSU's secure networks. It encrypts removable storage devices, such as USB sticks and external hard drives. It also prevents data from moving from OHSU's secure networks to unencrypted removable storage devices. Dell Data Protection includes a number of applications that detect data security risks on desktops, laptops, and external media; protect data on these devices by enforcing access control policies, authentication, and encryption of sensitive data; manage data centrally with policies using collaborative tools that integrate into existing user directories; support key and data recovery, automatic updates, and tracking for protected devices.

2. Procedures to restrict APAC data access to only those individuals listed on the data use agreement

   APAC data will be stored to an access-restricted folder within OHSU’s shared network drive. Access to the folder will only be granted to individuals listed on the DUA for the purposes of data management and analysis. All requests for remote access to the private network folder must first be approved by the data custodian, Nicole Weiskopf. Only a person with an active OHSU network ID access OHSU’s network. In addition, remote access to the restricted network folder requires two additional permissions. The user must have appropriate user permission to access the restricted network folder and they must be granted permission to Remote Desktop Protocol (RDP). RDP permission acts as a gatekeeper to prevent unauthorized access to the OHSU network. Users with access to this folder must be listed on the CPB study with OHSU’s IRB, which means they have up to date research ethics training and a completed a conflict of interest (COI) form. In order to add external collaborators from OCHIN. The data custodian will request

from the OHSU Department of Medical Informatics and Clinical Epidemiology (DMICE) IT Contact to

- request a network account (Temporary User Profile) for the non-employee collaborators at OCHIN who will access secure data for this project
- submit a request for RDP access for all users who need to access restricted network location
- revoke any of these permissions when access is no longer needed

Access to APAC data will be limited to CPB study members who are listed on the study IRB and the DUA which includes the data custodian (Nicole Weiskopf); OHSU faculty (Steven Bedrick, David Dorr, and Ana Quiñones); and University of North Carolina (UNC) Chapel Hill study members who are covered by OHSU IRB (Caroline Thompson); and CPB study members located at OCHIN (Teresa Schmidt, Matthew Jones, Lily Cook and Rae Crist). The UNC and OCHIN users will have remote access to this OHSU network location folder for limited duration (~6 months) to conduct data management procedures. If there are any changes to staff, we will revoke permissions for those who step down and complete the process described above for new staff.

3. User account controls, *i.e., password protections, maximum failed login attempts, lockout periods after idle time, user audit logs, etc*.

   All user accounts are centrally managed by OHSU IT. Password requirements include minimum length and complexity as well as regular expiration dates. OHSU has a standard 20 minute session time out. Maximum failed login attempts are 5, at which point users are lockout for 10 minutes before they may enter their password. Remote access requires two-factor authentication software. User activity is audited on the OHSU network.

4. Confirmation of training for personnel on how to properly manage protected health information in all formats

   All study personnel have up-to-date CITI certifications for Responsible Conduct of Research and Health Services Research. CPB analysts are trained to follow federal HIPAA regulations, which require specific protocols for the transferring, storage, and reporting of protected health information. Both OHSU And OCHIN conduct annual privacy and security trainings. OHSU employees are required to annually complete an updated “Information Privacy and Security Essentials” training. The training includes scored assessment, and employees must receive a passing score.

5. Protection of derivatives of APAC data at the re-identifiable level

- *If applicable, procedures for handling direct identifiers, such as allowing access on a 'need to know' basis only and minimizing risk by storing identifiers separately from other APAC data*

  To ensure minimum data is shared, the OCHIN team will provide a finder file to OHA staff, containing deidentified CPB Patient IDs alongside direct patient identifiers (e.g., name), so that OHA can develop a restricted, deidentified dataset of APAC claims data specific to CPB patients. The CPB Patient IDs will be randomly generated unique IDs that do not include any PHI or PII. OHA will use the direct patient identifiers to link CPB patient IDs to APAC person IDs and will then provide OHSU with (1) APAC claims for CPB patients who meet the inclusion criteria specific to this application, and (2) an CPB-to-APAC crosswalk linking deidentified CPB patient IDs to deidentified APAC person IDs. The crosswalk table that links CPB patient IDs and APAC person IDs will contain only randomly generated unique IDs and will not include any PHI or PII from OCHIN or OHA. Thus, OHSU will never receive personal identifiers on CPB patients from either OCHIN or from OHA.

  To define CPB outcome variables (e.g., ED visits, inpatient stays, and cardiovascular events), OCHIN staff will request remote desktop permission to access the OHSU-hosted APAC data. Matthew Jones and Lily Cook will create analytic tables that operationalize the variables of interest from APAC data, storing these analytic tables in the same restricted folder on OHSU's secure server. Additional deidentified datasets from OCHIN, which will include information obtained from OCHIN's research data warehouse, will also be added to this restricted folder on OHSU's secure server. OCHIN's limited datasets will contain the same deidentified CPB patient IDs that they provided to OHA to use in the CPB-to-APAC crosswalk. This deidentified CPB patient ID will allow analysts on the project to link APAC variables of interest with OCHIN variables of interest on the same CPB patients. No direct identifiers on CPB patients will be provided to OHSU in the APAC datasets, and no direct identifiers on CPB patients will be provided to OHSU in any datasets shared for the CPB study. Only analysts on the project will have access to the deidentified data for variable operationalization, for linkage between APAC and OCHIN tables, and for analysis.

6. Procedures for identifying, reporting and remedying any data breach

   OHSU requires any new or increased risk related to a study, including adverse or protocol deviation, be submitted to the IRB as Reportable New Information (RNI) within 5 business days. Potential breaches of confidentiality must be reported to both the Institutional Review Board and the Office for Information Privacy and Security as soon as possible. The IRB evaluates each reportable event submission to determine whether it meets the regulatory definition of an unanticipated problem involving risks to subjects or others and/or an instance of serious or continuing noncompliance.

7. Statement of compliance with HIPAA and the HITECH Act

As a covered entity, OHSU is subject to the HIPAA Privacy Rule including provisions that apply to research involving the use or disclosure of Protected Health Information (PHI).

8. Electronic device protections, i.e., anti-virus or anti-malware software, firewalls, and network encryption

All APAC data will be stored and backed up on a restricted network folder using secure servers. All OHSU computers are installed with the FireEye endpoint protection tool for antivirus and threat detection,and are encrypted with Bitlocker (Windows) or FileVault (MacOS). OHSU requires all restricted data at rest to be encrypted.

OCHIN workstations, laptops, and mobile devices attached to the network have their disks encrypted. Anti-virus and anti-malware is centrally-managed, and operating systems patches are applied through administrative controls without user intervention. The network has multiple layers of firewall protection and multiple firewall manufacturers.

b. Record level or derivative data that can be re-identified must be destroyed within 30 days of the end of the data use agreement, in a manner that renders it unusable, unreadable or indecipherable. What are your plans for destruction of the dataset and any potentially identifiable elements of the data once the data use agreement has expired?

At time of data destruction, study staff will reach out to OHSU IT staff to delete the restricted network folder and any previously backed-up versions located on the shared network drive.

7.2 Optional Racial Justice Addendum:

The overall purpose of this project is to improve the accuracy of clinical algorithms that are commonly used to identify at-risk patients, specifically with a goal of ensuring equity in their performance of these algorithms for patients of all racial and ethnic backgrounds, as well as across other social determinants of health (SDoH). This work is important because these algorithms are used to identify patients requiring additional care and resources, guide treatment and testing, and contribute to overall resource allocation. Prior research has found that electronic health record (EHR) data, which feed into these algorithms, are of lesser quality for Black and Hispanic patients compared to non-Hispanic White patients. While our project will not actually implement the improved algorithms in the clinical setting, we expect that it will serve as necessary foundation to future implementation, and that our findings will guide other research into the relationship between health equity, clinical data quality, clinical algorithms, and clinical care.

As described in section 1.2, our aims are to 1) determine if the performance of these algorithms varies with respect to race, language, and other SDoH; 2) explore the quality of clinical data these algorithms ingest and test if social determinants of health are drivers of clinical data quality; and 3) develop improved versions of these clinical algorithms to improve overall performance and mitigate racial, language, and other inequities that are identified in the first two aims. Our efforts will be centered on six validated clinical prediction and risk assessment tools that are commonly used in clinical settings to predict cardiovascular risk, all-cause mortality, and hospital readmissions.

APAC data serve two purposes in this project. First, the key outcomes that are predicted by the target algorithms may not be documented in our primary source of clinical data, which will be provided by OCHIN, but are likely to be present in APAC data. The majority of OCHIN data are from ambulatory healthcare settings, and therefore outcomes like cardiovascular events and hospital admissions are likely to be absent. These outcomes will serve as “labels” for our algorithm evaluation, model training, and model testing.

Second, because we believe that performance of these algorithms is partially dependent on the quality of the data to which they are applied, and that data quality partially mediates the relationship between patient SDoH and algorithm accuracy, assessing the quality of clinical data is a key focus of this project. The quality of some data elements can be assessed through simple metrics of data completeness or timeliness, but accuracy is challenging to determine without comparison to other sources of data. APAC data would provide an ideal opportunity to assess data accuracy by comparing or clinical data to patient claims, especially because APAC data are more robust against healthcare fragmentation than OCHIN data.

Our approach to this project is guided by causal inference methods and theory, which require that the relationships between relevant variables are informed by knowledge rather than data-driven associations. This approach will help us avoid many of the recent pitfalls around the inclusion of race, ethnicity, and other SDoH in healthcare and biomedical research. For example, we do not believe that race has a direct impact on cardiovascular health, but rather that experiences structural racism and implicit and explicit bias in

healthcare lead to different health burdens, healthcare quality, and healthcare outcomes. While we are not able to directly measure experiences of racism, a causal inference approach allows us to incorporate such concepts as latent factors, leading to more valid findings. This approach also allows us to identify and address possible intersections between race, ethnicity and other patient-level factors including, but not limited to, socioeconomic status, disability, and access to healthcare.

In short, we believe that this project aligns with the goals and priorities of the Oregon Health Authority focus on using APAC data to eliminate racial injustice and improve health equity. All findings will be shared with Oregon Health Authority and disseminated broadly via presentations and publications (we will be paying for open access fees as needed). We will also be working with a clinical advisory group to identify potential avenues for implementation going forward. Research output and methods, including the improved algorithms, will also be made freely available. Our commitment to sharing findings and research products is written into the NIH research proposal that is funding this work.

Minimal Risk Protocol Template

1) Protocol Title

Health equity and the impacts of EHR data bias associated with social determinants

2) Objectives

Our long-term goal is to improve the performance of clinical risk prediction tools, thereby enabling appropriate clinical actions for all patients, especially those with adverse SDoH. This goal requires both data-driven elucidation of the mechanisms driving data bias and biased predictive accuracy, and the application of novel methods to mitigate these biases. Based on existing research and our own preliminary data, we hypothesize that adverse SDoH are drivers of poor EHR data quality, and that EHR data quality in turn drives clinical risk prediction validity, leading to estimates of risk that are more valid for some patient populations than others. The rationale for our approach is that by understanding how SDoH impact data quality and risk prediction, we can apply established debiasing methods to develop augmented clinical risk prediction tools that are robust against bias. We will leverage a unique, largescale dataset from a community-based health research network that serves a representative patient population with high-quality SDoH data, linked to reliable outcome data that will serve as ground truth. We propose the following three aims:

SA 1: Implement patient-level data quality assessments to characterize relationship between EHR data quality and SDoH. We will develop and apply EHR data quality metrics to a large dataset from a community-based healthcare research network with enriched SDoH data. We hypothesize that adverse SDoH lead to worse quality EHR data, mediated by healthcare utilization and health status.

SA 2: Determine impact of adverse SDoH and EHR data quality and on clinical risk prediction validity. We will measure clinical prediction validity by retrospectively comparing predicted and actual outcomes for six widely-adopted clinical risk prediction algorithms. We hypothesize adverse SDoH lead to lower risk prediction accuracy, and that this relationship is mediated by EHR data quality.

SA 3: Apply debiasing methods to create augmented risk prediction algorithms that are robust against adverse SDoH and poor data quality. We will build structural models of the data generating mechanisms for each selected clinical risk prediction tool, which will inform the application of debiasing techniques to create augmented versions of the risk prediction tools. We will test these models for mitigation of bias and improvements in predictive accuracy, stratified by SDoH.

Our expected outcomes are 1) characterization of mechanisms linking SDoH, EHR data quality, and algorithm-supported risk prediction accuracy; 2) the generation of debiased risk prediction algorithms that are robust against data quality problems driven by adverse SDoH; and 3) identification of patient characteristics whose presence indicate that risk prediction tools may underperform.

3) Background

A1. Reduction of health disparities is one of the biggest priorities in improving American health.

Within the U.S., substantial differences in mortality rates, longevity, infant mortality, and other common metrics of health exist along racial and ethnic lines.[15-17] Race and ethnicity are not, however, the causes of these disparities. Rather, they are proxies for systemic racism, other forms of discrimination, and socioeconomic status,[18-20] all of which are associated with negative health experiences and outcomes. As of 2017, the CDC reported that age-adjusted mortality rates were

13% lower for non-Hispanic adult whites than for non-Hispanic adult Blacks.[21] Similarly, CDC data on infant mortality from 2018 showed rates of infant death more than double for non-Hispanic Black women compared to non-Hispanic whites.[22] These disparities exist in specific clinical contexts as well. For example, COVID-19 has disproportionately impacted historically marginalized racial and ethnic groups, individuals with financial or housing insecurity, and individuals otherwise considered to be at-risk or vulnerable.[23] COVID-19 disparities are observed in testing rates,[24] case rates, hospitalizations, and mortality.[25] To successfully reduce health disparities, we must consistently identify and act on them.

A2. Clinical risk prediction tools are important for ensuring high-quality care, but may not be equally valid for all patient groups, and may exacerbate disparities. The identification of clinically at-risk patients is of high importance in healthcare and care management. Risk assessment models and clinical prediction models may be used to determine whether a patient needs intervention, including specific medical treatments, lifestyle changes and counseling, and resource allocation in the medical setting. Patients determined to be at high risk for cardiovascular disease, for example, may be started on statin or aspirin therapy, encouraged to adopt lifestyle changes; and undergo additional medical testing and assessment (e.g., additional stratification through imaging or stress testing).[26, 27] Each such intervention requires additional time and effort from providers and patients, has associated costs, and may bring risks and side effects. For the right patients, though, these interventions reduce the risk of undesirable medical outcomes, improve quality of life, and decrease longer term medical costs.[28-30] It is therefore of great importance that the EHR algorithms used to identify at-risk patients are valid and reliable.

These algorithms, however, do not perform equally well for all patients. The Framingham Coronary Heart Prediction Score, for example, was found to overestimate risk for Japanese–American men, Hispanic men, and Native American women unless calibrated using data from cohorts more representative of those racial and ethnic groups.[31-33] More recently, Obermeyer et al. demonstrated that an algorithm used to estimate healthcare costs was more likely to systematically underestimate future utilization and costs of Black patients, leading to fewer resource allocated and worse outcomes for this population.[2] Similarly, work by Taylor et al. suggests racial bias in MEWS and NEWS scores for predicting in-hospital mortality.[34]

Additionally, many risk scores and algorithms intentionally include race and ethnicity as adjustment variables without appropriate evidence or consideration of societal context. Vyas et al. reviewed several such tools and highlighted the fact that many of them automatically downgrade severity for Black patients, potentially depriving patients of clinical interventions that would have been made available had they not been Black. They emphasize the necessity of "reconsidering race correction in order to ensure that our clinical practices do not perpetuate the very inequities we aim to repair."[1] Similarly, there is a growing literature on the importance of considering how we treat race, ethnicity, and other SDoH in the development, evaluation, and application of machine learning and data mining methods,[35-37] including those that focus on genomic data.[38] All of this work emphasizes that race, ethnicity, and other SDoH must be treated not as clinical variables, but rather as indicators of societal and systemic experiences of racism, implicit bias, and limited access to care.

A3. Clinical risk prediction errors cannot be entirely attributed to faults in the algorithms themselves; the quality of the data to which they are applied must be considered. To the extent that adverse SDoH are associated with risk prediction failures, it is vital that we determine whether these failures are due to faulty models or problems with the data to which the models are applied.

EHR data are largely a byproduct of complex processes within the healthcare system related to billing and the documentation of immediate medical concerns.[39] It is not surprising, therefore, that research has found EHR data quality to be highly variable and often insufficient.[8-10] EHR data quality problems include, but are not limited to: data that are implausible, incomplete, out-of-date, and non-conforming (not properly formatted or adhering to standards).[40, 41] [40, 41]

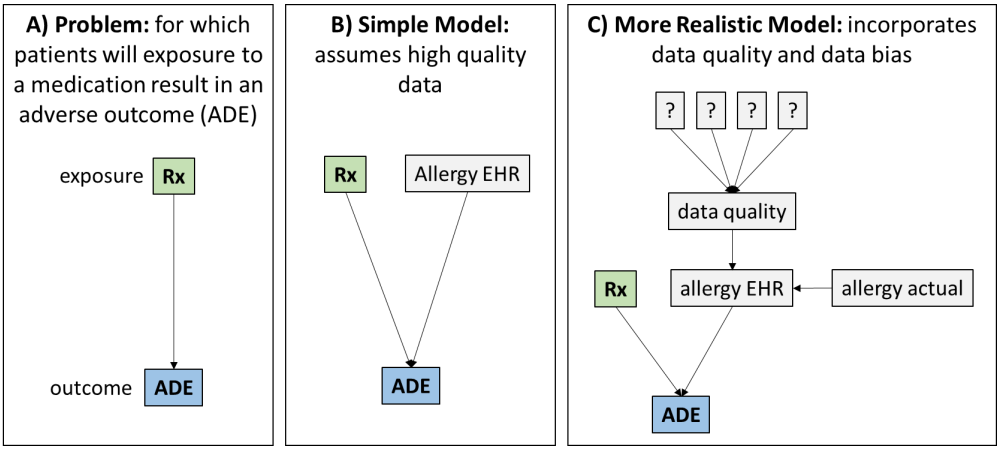

Figure 1: Structural models expressing relationship between a medication (Rx) exposure and an adverse drug event (ADE) outcome. A) If we can predict which patients are likely to have an ADE from exposure to Rx, providers intervene to prevent that outcome. B) If medication allergy is recorded, clinical decision support can be used to predict an ADE and trigger an alert. C) Observed allergy in EHR is a descendent of actual allergy status *and* data quality.

To demonstrate the potential impact of applying a rule-based algorithm to poor quality clinical data, we offer the following example. EHR systems typically include automatic alerts warning of potential medication-related errors. Consider an alert intended to warn prescribing providers that a patient is likely to have an an allergic reaction from exposure to a medication. This example is illustrated in Figure 1. If medication allergy information is recorded, clinical decision support can be used to predict an allergic reaction and trigger an alert, preventing the ordering of that medication (Figure 1, panel B).

The clinical decision support system, however, doesn't know if the patient *actually* has an allergy, only if an allergy was recorded in a structured EHR field. Observed allergy status from the EHR is not only the descendent of true allergy status ("actual allergy"), but its validity is also affected data quality problems.

A4. EHR data quality problems do not occur at random, and are likely driven at least in part by SDoH. While general EHR data quality has received significant attention over the last several years, with a number of proposed conceptual frameworks and tools,[40, 42, 43] systematic biases in EHR data (i.e., data quality problems that do not occur at random) are less understood. Our prior research, which has been confirmed by others, shows that sicker patients have a greater quantity of data in their records than healthier patients.[11, 12, 44, 45] This makes intuitive sense and is perhaps clinically appropriate, but will impact the validity tasks that rely on these data. If we return to Figure 1, panel C, and assume that one of the drivers of data quality (represented by question marks) is patient health status, it follows that allergy data in the EHR are likely to be more complete and accurate for sicker patients. This has the potential to result in a higher error rate for healthier patients.
If data quality is driven not only by health status, but also by SDoH, then performance of this allergy alert would also be biased with respect to SDoH.

Research on the quality of EHR data as a function of SDoH is limited, but published research supports our hypothesis that EHR data quality is worse for patients with adverse SDoH. Klinger et al. showed that the sensitivity of structured race and ethnicity EHR data was higher for non-Hispanic whites than for Blacks and Hispanics.[14] Sholle et al. found that Black and Hispanic patients without race and ethnicity data were older, sicker, more likely to be male, and less likely to have commercial insurance.[13] In the medication allergy example above, if the quality of data for allergy is also driven by SDoH, the presence of unreliable data for underserved patient populations could exacerbate inequities in care.

More comprehensive and systematic research on this issue is required. Published literature and our own work on this topic have touched on a limited number of clinical variables, data quality metrics, and SDoH. Additionally, key covariates relating to health status and healthcare utilization must be included in analyses in order to truly characterize the role of adverse SDoH in driving EHR data quality. We also note that the use of OCHIN data provides a unique opportunity to study this phenomenon. The EHR-derived SDoH data most often studied are often of poor quality,[46] and are reflective of patient populations treated at academic medical centers. The SDoH data captured by the OCHIN network are both high-quality and representative of underserved populations. One of the primary deliverables of the proposed work is an in-depth, systematic exploration of the association between different SDoH and EHR data quality in a representative patient population. This is responsive to objective one of the National Library of Medicine Notice of Special Interest NOT-LM-23-001.

Our preliminary findings also indicate that EHR data quality varies with respect to SDoH. For a sample of 235 cardiology patients at OHSU, we determined whether the structured data recorded in the EHR for seven diagnoses and five medication classes associated with cardiovascular care were true positives, false positives, true negatives, or false negatives compared to a reference standard derived from manual chart review and patient self-report. The average number of data points that were missing or incorrect for each patient was 1.9 out of 13 (14.5%) and the median was 2 (15.4%). A mixed-effect model was fitted to the data, with a set of commonly-recorded SDoH as the independent variables. We found that financial stress, male sex, and increased age were significantly associated with the likelihood of a false positive or false negative (**Figure 2**).

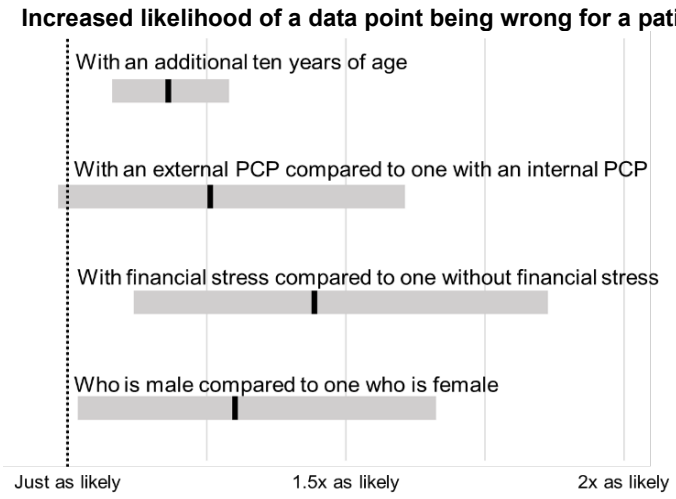

Figure 2. Increased age, financial stress, and male sex are associated with increased likelihood of clinical data being missing or incorrect.

The hypothesis that SDoH drives EHR data quality is also supported by preliminary analysis of data from 1.8 million records from OCHIN (the health network that will provide data for this proposal; see section C2b). We found that patients without insurance—an important SDoH—were in most cases less likely than insured patients to have a problem list history, basic vitals, or a selection of important labs (**Table 1**).

Table 1. Presence of common clinical concepts by insurance status for OCHIN patients aged 18-89 seen from 2018-2020. HT/WT: height/weight, BP: blood pressure, A1C: Hemoglobin A1c, INR: International normalized ratio, LDL: Low-density lipoprotein, HGB: Hemoglobin, CK: Creatine kinase, TROP: Troponin I cardiac or Troponin T cardiac.

	Medicaid	Medicare	Other	Private	Uninsured
Total Patients	**760,020**	**222,472**	**31,792**	**339,441**	**495,546**
Problem List History	86.4%	91.8%	87.8%	82.0%	75.5%
Any HT/WT Measurements	86.3%	89.4%	92.5%	83.0%	76.3%
Any BP Measurements	91.3%	91.1%	93.8%	86.8%	82.5%
Any A1c[1] Lab days	35.9%	51.1%	56.2%	32.6%	29.0%
Any INR[2] lab days	2.1%	4.2%	1.5%	1.1%	1.2%
Any LDL[3] lab days	37.4%	56.6%	55.7%	38.4%	29.6%
Any HGB[4] lab days	43.8%	53.7%	52.9%	38.1%	31.2%
Any Creatinine lab days	43.8%	65.2%	58.3%	43.6%	34.4%
Any CK[5] lab days	0.9%	1.8%	1.5%	0.7%	0.6%
Any TROP[6] lab days	0.1%	0.3%	0.1%	0.1%	0.1%

A7. Methodological approaches to ensure appropriate risk assessment and risk prediction despite poor quality EHR data are necessary. Identifying and developing methods that can be used to debias risk prediction and associated clinical practice is a priority,[47] and is directly responsive to objectives one and two of the National Library of Medicine Notice of Special Interest NOT-LM-23-001. Some forms of bias in EHR data may be addressed through the use of subpopulation calibration,[48] propensity scores, and other matching and weighting techniques,[49, 50] but these methods are often applied without a systematic understanding of how the features used to inform matching are associated with selection bias, information bias, or outcome validity. Therefore, the final deliverable of this proposal will be the application and evaluation of structural modeling approaches that incorporate SDoH concepts into our clinical risk prediction use cases, allowing us to control for and correct biases present in EHR data, as well as identify indicators that these risk tools are likely to fail for a patient, thereby allowing appropriate clinical intervention and action at the point of care.

A8. Expected impact.

While the work described in this proposal focuses on clinical risk prediction, the deliverables of this work are expected to have a broader impact. EHR data are also used for quality improvement efforts, cohort, and retrospective research. All these use cases would also be potentially impacted by data bias associated with SDoH. We propose a conceptual model (Figure 3) that links SDoH to healthcare utilization and health status, which in turn impact the quality of EHR data and, ultimately, the validity of EHR data reuse.

This research will provide novel and actionable information about bias in EHR data, disparities in clinical risk prediction accuracy, and methodological approaches to lessen the impact of these biases and disparities. **A systematic exploration of the association between EHR data quality and SDoH has not previously been conducted. Analysis of the impact of EHR data quality on clinical risk prediction validity is similarly novel.** The expected contributions and deliverables from this work will address important gaps in knowledge and practice individually and as a high-level framework. Specifically, we intend to address the following gaps:

1. Methodological approaches for assessing EHR data quality rarely address the topic of non-random data quality problems (i.e., bias). Moreover, the patient-level factors associated with variations in EHR data quality are largely uncharacterized.

→ *We will systematically measure EHR data quality and test for association with patient SDoH. These results will have applications in all forms of EHR data reuse.*

2. While disparities in clinical risk prediction with respect to SDoH like race and ethnicity have been documented, these failures have been largely attributed to faulty algorithm specifications or problems with the data used to develop them. The role of EHR data quality at the point of care has not been explored.
→ We will quantify the accuracy of clinical risk prediction using EHRs alone, vs. with augmentation with data from all-payer claims and death certificates (gold standard), and w*e will determine if and to what extent disparities in clinical risk prediction accuracy can be explained by variation in EHR data quality.*

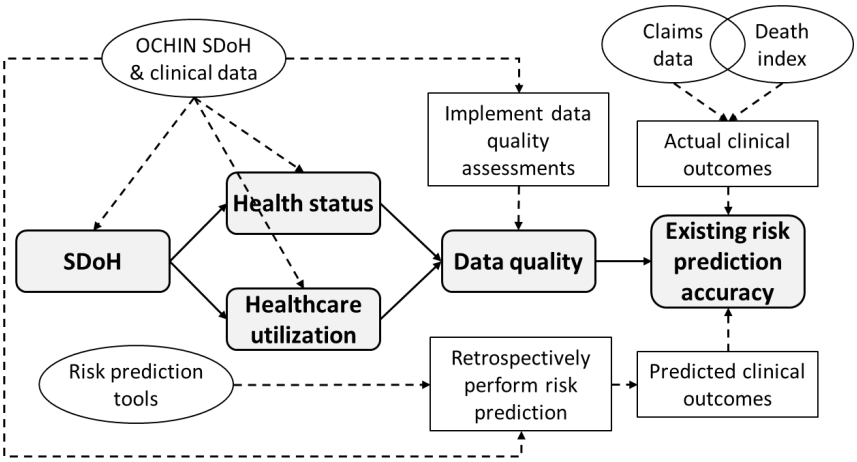

Figure 3. Conceptual model summarizing hypothesized relationships between SDoH, health-related factors, EHR data quality, and risk prediction accuracy. Key resources are depicted in ovals. The relationships between these concepts will be learned in Aims 1 and 2 and used to inform debiasing methods in Aim 3.

3. At present, the majority of work on the appropriate integration of SDoH into clinical algorithms is theoretical or knowledge-driven. Data-driven approaches are needed to identify and estimate biases so that improvements to clinical algorithms can be empirically grounded and tested for efficacy.
→ *We will use epidemiological methods to create graphical models reflecting the data generating mechanisms of risk prediction use cases, incorporating the concepts and relationships identified as meaningful in the first two aims. These models will be used, with our empirically derived parameter estimates from a gold standard population to de-bias the risk prediction tools and tested for their accuracy.*
4. There are substantial disparities in health and healthcare. Some of these disparities are furthered by clinical risk prediction tools that may over- or underestimate risk for at-risk and underserved patient populations.
→ *The methods developed in this work will promote awareness of how SDoH impact healthcare data quality, and improve validity of risk prediction allowing providers to identify patients for whom they may underperform at the point of care, additional interventions and/or information-gathering.*

4) Study Design

This will be a fully retrospective study, relying on existing clinical data. There will be no intervention and no patient/participant contact of any kind.

Our final deliverable is a set of debiased risk scores that have improved validity for all patients, regardless of adverse SDoH and problems with the EHR data used to calculate these risk scores. This work will be conducted using upon six existing, widely-adopted clinical risk prediction tools as use cases (see Table 3 below). Debiasing will be accomplished via the application of established methodology informed by structural models of the predictors of risk score accuracy. It is our hypothesis that at the patient level, the primary driver of risk score accuracy is the quality of the EHR data to which they are applied, which is in turn driven by SDoH, mediated by health status and healthcare utilization. We will test this hypothesis and determine the strength of the relationships between SDoH, health status, healthcare utilization, EHR data quality, and risk prediction validity. These learned relationships will be represented in the final structural models, which will then be used to generate the debiased risk prediction models.99 These models will then be evaluated for overall accuracy and compared to the original models.

Patient-level SDoH, health status, and healthcare utilization will be extracted or derived from our clinical dataset. Patient-level EHR data quality will be determined by applying existing or novel data quality

metrics to the clinical dataset. Risk prediction validity will be determined for each patient and for each relevant risk score by applying the existing algorithms to clinical data extracted from our dataset and comparing predicted outcomes to actual outcomes. Actual outcomes will be triangulated from our clinical dataset, all-payor claims data, and Social Security Death Index data.

5) Study Population

a) Number of Subjects

All of our data will be from OCHIN, who have agreed to waive IRB oversight to OHSU, and the Oregon Health Authority (OHA) All Payers All Claims (APAC) database. The first two aims will focus on data from adult patients with SDoH data present. At this time that includes approximately 300,000 patients. We do not expect that data from all of these patients will be used, but that would be the maximum. An additional 10,000 patients without enhanced SDoH data will be selected for model testing in aim 3.

b) Inclusion and Exclusion Criteria

OCHIN Criteria:

Eligibility will be determined based on data present in OCHIN EHR. No screening will be conducted.

Inclusion in all aims:

- Adult (18-89)
- Enhanced SDoH screening conducted

Inclusion in aim 3 for algorithm testing:

- Adult (18-89)
- Enhanced SDoH screening NOT conducted
- Meet eligibility criteria for at least one of the six risk prediction tools that will be used
    - Atherosclerotic Cardiovascular Disease risk calculator
    - Simple Framingham Risk Score (cardiovascular)
    - Charlson Comorbidity Index (10 year mortality)
    - Elixhauser Comorbidity measure (10 year mortality)
    - HOSPITAL readmission score
    - LACE readmission score

We do not have any specific exclusion criteria.

OHA APAC Criteria:

Inclusion:

- Adult (18-89)
- Meet any of the following criteria:
    - At least one diagnosis from the list of diagnosis codes used to calculate the Charlson Comorbidity Index or the Elixhauser Comorbidity Measure between 2011-2022
    - At least one prescription for statins, aspirin, or antihypertensives between 2011-2022
    - Patients with an inpatient admission or emergency department visit between 2019-2022
- Can be linked to an OCHIN record for a patient meeting the OCHIN inclusion criteria above

As noted above, we will not be conducting screening, so there will not be any disposition of screening data.

c) Vulnerable Populations

Only adult participants will be included in this study. We will make no effort to include or exclude members of other vulnerable populations, and it is possible that members of such populations may be incidentally included. We do not anticipate any undue risk to such participants.

d) Setting

The primary site will be OHSU. Other sites are OCHIN and University of North Carolina, Chapel Hill. Both OCHIN and UNC will waive IRB oversight to OHSU.

OCHIN: All EHR-derived data will be provided by OCHIN. The OCHIN personnel will be responsible for extraction, transformation, and transfer or relevant data to the primary site. We do not expect that analyses will be conducted at OCHIN, though OCHIN personnel will help guide analyses.

University of North Carolina: UNC will not provide any data. They will help with guiding analyses in the first two aims, and will be responsible for the bulk of the analyses conducted in Aim 3.

OHSU: OHSU will not provide any data, but will be responsible for storage and analysis of data after they have been securely transferred from OCHIN to OHSU secure storage.

e) Recruitment Methods

This will be a fully retrospective study, so no recruitment will be conducted.

f) Consent Process

Because this is a fully retrospective study, we will not be seeking consent from patients whose data we use. We will instead seek a waiver of consent.

6) Procedures

This is a fully retrospective study with no patient/participant interaction of any kind.

7) Data and Specimens

a) Handling of Data and Specimens

Data will be limited to structured clinical data from the OCHIN EHR network, Social Security death index data, and, for patients residing in Oregon, linked claims data from OHA APAC. These data will include demographics, social determinants of health, visit and encounter data (e.g., dates and encounter types), diagnoses, laboratory results, vitals, orders, medications, and death information.

After extracting and transforming these data as necessary, OCHIN will securely transfer the data to OHSU, where they will be securely stored. We expect that these data will reside on OHSU's OneDrive instance, the secure network storage provided by ACC, or on OHSU managed and encrypted computers. Additional data from OHA APAC will be linked with OCHIN data at the patient level by a third-party vendor contracted by OHA. Linked APAC data will be transferred directly to OHSU.

OHA APAC data are requested via a detailed application that requires information about the study, scientific knowledge to be gained, and potential impact on patient care and health. Additional information regarding data security and human subjects protections must also be provided. Our data request to APAC is as follows:

The data request submitted to OHA APAC will be as follows.

> Specifically, we'd like to request APAC claims data for study patients who meet any of the following criteria, and a list of study patients who do not meet any of these criteria:
>
> 1. Please provide all claims 2011-2022 for patients with any diagnosis from a list of diagnosis codes that we will provide (for the Charlson Comorbidity Index and Elixhauser Comorbidity Measure).
>
> 2. Please provide all claims 2011-2022 for patients with any prescription from a list of relevant drugs we will provide NDC codes for (including statins, aspirin, and HTN medications).
>
> 3. Please provide claims 2019-2022 for patients with claims to indicate an inpatient admission or an ED visit, based on APAC groupers for defining inpatient visits and ED visits
>
> We would also like to request that the patient crosswalk (from OCHIN Client_ID to APAC member_ID) include member_IDs for patients who did not meet any eligible criteria. That will help us to differentiate which of our OCHIN study patients did not link to any APAC member IDs, versus those who linked to someone but did not meet any of the above inclusion criteria.

For the third aim, relevant data (which will have undergone further transformation, standardization, and deidentification) will be securely transferred to University of North Carolina at Chapel Hill.

b) Sharing of Results with Subjects

No findings will be shared with patients.

c) Data and Specimen Banking

Data will not be retained after the conclusion of the study and related analyses. We will not be creating a repository, and data will not be shared or made available.

8) Data Analysis

C3. SA 1: Implement patient-level data quality assessments to characterize relationship between EHR data quality and SDoH. The purpose of this aim is to broadly and systematically measure the impact of adverse SDoH on EHR data quality for a patient population that is representative of historically underserved groups. As noted in the Significance section, existing literature and our preliminary data support the existence of this relationship, but *this prior work is scarce and has been applied to a limited set of SDoH variables and data quality assessments, mostly based on patients seen at academic medical centers*. We will also determine if and to what extent the association between SDoH and EHR data quality is mediated by healthcare utilization.

We will rely on two conceptual frameworks to guide our measurement of data quality. The Harmonized Data Quality Assessment Framework, developed by multiple experts in the field, will be complemented by 3x3 DQA, developed by Dr. Weiskopf.[40, 42] The Harmonized framework includes assessment approaches for completeness, conformance (adherence to standards), and plausibility (accuracy). 3x3 DQA has the additional data quality category of currency (timeliness or "ripeness" of data) and provides a more in-depth approach to completeness assessments.[62] Overall data quality will be assessed via measurements of data completeness and use case-specific data quality (i.e. sufficient completeness, plausibility, conformance, and currency) will be assessed via data quality checks.

Hypothesis: Patient-level EHR data quality is worse for patients with adverse SDoH.

C3a. Data processing and variable generation

As described above, data for this aim will be limited to OCHIN clinics with an SDoH screening rate of at least 85% (**C2bii**).[100, 102] We will include all patients seen during the measurement period at these clinics in order to avoid the pitfalls of complete case analysis. We anticipate that the data quality assessment portion of this aim will be the most time- and resource-intensive part. Data quality is typically defined as intrinsic (i.e., objective, inherent to the data) or extrinsic (i.e., subjective, fitness for use).[109, 110] Therefore we will calculate multiple metrics of data quality. Intrinsic data quality will be operationalized as overall completeness, or the quantity of clinical data present. Extrinsic data quality will be operationalized with respect to patients' clinical status as well as relevant use cases (risk score calculation). These data quality metrics will be calculated for each patient, not for the overall dataset.

C3ai. Dependent variable: Overall completeness: As demonstrated in Dr. Weiskopf's previous research, overall record completeness will be operationalized as counts of basic clinical data types that are present, current, and conformant,[62] yielding an count of data points for each patient. To summarize:

overall completeness for a patient = the count of medication orders, laboratory results, diagnoses (problem list or encounter), vitals, and progress notes documented in the patient's record during the study period that are conformant to relevant standards

C3aii. Dependent variable: Data quality defined by use case: Use case-specific data quality is defined as fitness-for-use: data are of good quality if they are sufficient for their intended purpose(s).[62, 110] We will calculate an overall data quality score for each patient record with respect to the six use cases, which will be the proportion of the 59 unique concepts (**Table 3)** that are complete, conformant, current, and plausible.

Completeness, conformance, and currency are straightforward to assess for concepts that are required in order to calculate a risk score at all.[40] See section **C4ai** on data sufficiency below for a list of these concepts. If one of the concepts required for calculation is not present in the EHR in a format suitable for calculation, or was not documented within the required time span, then that concept does not meet the requirements for task-dependent data quality. As an example, to calculate an ASCVD risk score one must be able to extract a systolic blood pressure value that adheres to relevant LOINC standards and appropriate units of measurement, and which was measured and documented within the past twelve months.

For concepts that are only present when relevant, such as diagnoses or medications, the approach to data quality assessment is more complicated. Returning to ASCVD as an example, if a diabetes mellitus diagnosis is present in a patient's record, do the associated concepts (e.g., glucose values, A1c values, insulin order, etc.) support the diagnosis? And alternatively, if the diagnosis is not present, are there associated concepts recorded that suggest it should be? Relevant plausibility checks of this variety will be drawn from existing sets of phenotypes and data quality assessments, including PheKB, OHDSI Data Quality Dashboard, OHDSI Phenotype Library, and Callahan's review of existing data quality checks. [89, 111-113] In summary:

use case-specific data quality for a patient = (count of distinct clinical concepts used in one or more of the risk prediction tools that are complete (correctly absent or correctly present), conformant to relevant standards, current, and plausible) / (total unique concepts)

C3aiii. Dependent variable: Data quality defined by clinical status: Because EHR data quality, especially completeness and currency, are known to be driven by patient health,[11, 12, 44] it is important to consider clinical status when assessing data quality. Unfortunately, indicators of overall clinical status that aren't themselves dependent upon the present of high quality data are available only for a minority of patients. Therefore, we will use a condition-based approach to measuring this form of data quality. For a given patient we will determine the presence of relevant conditions by considering problem list and encounter diagnoses. For each condition we will check for the presence of additional concepts (e.g., laboratory results and medications) that are expected to be present for those diagnoses. These expected concepts will be determined by reviewing best-practice guidelines and consultation with our advisory panel. The resulting metric will be the proportion of clinically-relevant concepts that are present for a given patient. The denominator for this metric will be defined by clinical status.

As a result, a patient with very little data but no diagnoses may end up with a higher result than a patient with a great deal of data and many diagnoses.

clinically-defined data quality for a patient = (count of all expected concepts for each documented condition that are present, conformant to relevant standards, current, and plausible) / (count of all expected concepts for each documented condition that are present)

C3aiv: Independent variables: SDoH and utilization: There are seventeen patient-level SDoH concepts, as defined above in section **C2bii**, which can be grouped into seven SDoH domains. These data may be enriched with geocoded community-level SDoH data as appropriate (**C2biii**). Utilization estimates will be derived from a combination of OCHIN EHR data and linked claims data where available (**C2biii**).

C3b: Analysis approach

We will apply appropriate statistical methods to test for and quantify relationships between SDoH and EHR data quality, as well as to determine which SDoH are significant and meaningful in predicting data quality.

C3b1. Statistical approach: One model will be learned for each data quality metric. Our first choice of statistical model is regression, which has the advantage of providing interpretable coefficient estimates that can be easily applied in Aim 3. The choice of regression model will depend on the data, but we anticipate, based on our past research,[62] that our data will follow a Poisson or negative binomial distribution. If necessary, non-parametric alternatives will be considered. We will consult with the OHSU Biostatistics & Design Program to ensure appropriate model selection and specification.

Given the nature of the patient factors that we will be collecting, it very likely that we will observe collinearity between our independent variables. We will perform appropriate tests of collinearity and, depending upon findings remove redundant variables or combine closely correlated variables (e.g., collapsing engagement and literacy into one construct: medical activation). We will also test for interactions and add interaction terms to our model as needed.

We believe that healthcare utilization is a mediator of the relationship between SDoH and data quality and SDoH. Based on the steps outlined by Kenny et al.,[114-116] we will run additional tests with utilization metrics to determine: 1) if utilization is dependent on SDoH, 2) if completeness is dependent on utilization, and 3) if utilization is a significant contributor to the association between SDoH and completeness.

C3bi: Power and sample size considerations: Between the SDoH and utilization variables described above (**C3aiii**), we anticipate a *maximum* of 20 independent variables. A regression model with an alpha of 0.01, beta of 0.95, a small effect size ($f^2 = 0.02$) and 20 predictors requires a sample of approximately 2,000 patients, which is well below the 8,800 patients from the SDoH OCHIN clinics (**C2bii**).[117]

C4. SA 2: Determine impact of adverse SDoH and EHR data quality and on clinical risk prediction validity. In this aim we will determine if the data biases explored in Aim 1 (data quality driven by SDoH) have an impact on the accuracy of clinical risk prediction. Clinical prediction validity will be measured by retrospectively comparing predicted and actual outcomes for the clinical prediction use cases. Associations between SDoH and quality of risk prediction will be tested, as will associations between clinical prediction validity and data quality. The most time- and resource-intensive part of this aim will be the retrospective calculation of the use case risk scores using OCHIN clinical data. The ideal outcome of this aim is that clinical risk prediction accuracy can be inferred from adverse SDoH without including data quality metrics as independent variables, since data quality is hard to assess, especially in real-time.

H1: Calculated risk scores are significant predictors of actual clinical outcomes.
H2**:** Patients with worse quality data are less likely to have accurate risk prediction scores.
H3**:** Patients with adverse SDoH are less likely to have accurate risk prediction scores.

C4a. Data processing and variable generation

As described above, data for this aim will be limited to OCHIN clinics with an SDoH screening rate of at least 85% (**C2bii**).[100, 102] We will include all patients seen during the measurement period at these clinics in order to avoid the pitfall of complete case analysis.

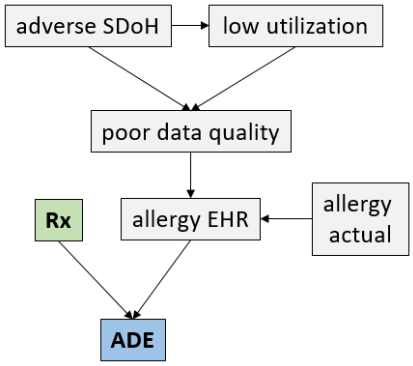

Figure 4. Revisiting DAG from Fig 1. ADE: adverse drug event, Rx: medication. Allergy in EHR is a collider for true allergy and poor data quality.

C4bi. Independent variables: SDoH: Extracted as described above in **C3aiv**.
C4bii. Independent variables: Data quality metrics: Calculated as described above in **C3ai**, **C3aii**, and **C3aiii**.
C4biii. Independent variables: Risk scores: *Predicted* outcomes or risk scores will be calculated for each eligible patient for each model. All risk scores will be calculated using OCHIN clinical data.

HOSPITAL and LACE readmission scores will be calculated for patients who were admitted as inpatients.[103, 104] Patients with multiple inpatient stays will have multiple scores calculated. While OCHIN clinics almost exclusively provide outpatient care, information about emergent and inpatient encounters at partner organizations and providers will often be linked to OCHIN records. Where available, claims data on hospital encounters will be used to augment EHR data (**C2biii**).

The ASCVD risk score will be calculated for patients who were between 40 and 75 years of age and had not developed cardiovascular disease or experienced a cardiovascular diagnosis upon their first presentation to the OCHIN medical system within our measurement period.

The "bedside" Framingham risk score risk score will be calculated for patients who were between 20 and 79 years of age and had not developed cardiovascular disease or experienced a cardiovascular diagnosis upon their first presentation to the OCHIN medical system within our measurement period.

The Charlson and Elixhauser Comorbidity scores will be calculated for all patients in our sample.

C4aiv. Dependent variables: Clinical outcomes: ***Actual*** clinical outcomes will be extracted from structured EHR data, the Social Security Death Index, and OHA APAC claims data for each patient. For each patient and tool (and each inpatient admission, in the case of HOSPITAL and LACE), the outcome variable will be dichotomous: the outcome of interest either occurred or did not occur in the specified time period.

HOSPITAL and LACE: The outcome for both tools is readmission or death within 30 days.[103, 104] Readmission data will be extracted from OCHIN EHR data and OHA APAC claims data. Data on deaths will be extracted from the Social Security Master Death File.

ASCVD and "bedside" Framingham: Using OCHIN EHR data and OHA APAC claims data, we will determine if each eligible patient developed cardiovascular disease or experienced a cardiovascular event (e.g. stroke or myocardial infarction) within five years and within ten years (if sufficient data are available) of the baseline visit defined above.[105, 106]

Charlson Comorbidity Index and Elixhauser Comorbidity Score: Using data from the Social Security Death Index, we will determine will five- and ten-year-mortality from baseline for patients.[107, 108]

C4b. Analysis approach

To address hypothesis 1 above and ensure basic validity of our risk score calculations, for each risk score we will test if the risk scores (**C4aii**) are meaningful and significant predictors of relevant clinical outcomes (**C4aiii**). We will also generate receiver operator characteristic curves for each risk score. If we do not find significant associations between risk scores and outcomes as described in the literature, we will reexamine our data extraction and transformation processes.

Hypothesis 2 will be tested by adding data quality metrics to the models used to test hypothesis 1. The dependent variable will still be observed outcomes, and the independent variables will include data quality metrics in addition to the calculated risks. This is essentially a specific application of mediation analysis, which Mittlbock and Schemper describe as explaining variation in predictive accuracy.[118, 119] If data quality metrics are

significant in these models, it means that variation in predictive accuracy is partially explained by patient-level data quality.

Hypothesis 3 will be tested in the same fashion as Hypothesis 2, except instead of data quality metrics we will be adding SDoH and metrics of health status and utilization to the independent variables. If SDoH are significant in these models, it means that variation in predictive accuracy is partially explained by SDoH, indicating bias in risk score predictive accuracy.

C4b1. Statistical approach: For all analyses described above, our first choice of statistical test is regression, which has the advantage of providing interpretable coefficient estimates that can be easily applied in Aim 3. The choice of regression model will depend on the data, with logistic regression being the obvious choice for dichotomous outcomes. If necessary, non-parametric alternatives will be considered. We will consult with the OHSU Biostatistics & Design Program to ensure appropriate model selection and specification.

C4b2. Power and sample size considerations: Between the SDoH and utilization variables described above (**C3aiii**), we anticipate a *maximum* of 20 independent variables for hypotheses 2a, 2b, 3a, and 3b. A regression model with an alpha of 0.01, beta of 0.95, a small effect size ($f^2 = 0.02$) and 20 predictors requires a sample of approximately 2,000 patients, which is well below the 8,800 patients from the SDoH OCHIN clinics (**C2bii**).[117] Hypothesis 1 will have only a single predictor, and will therefore also have sufficient sample size.

C5. SA 3: Apply debiasing methods to create augmented risk prediction algorithms that are robust against adverse SDoH and poor data quality. We will use the findings from Aims 1 and 2 to reduce bias and improve the accuracy of clinical risk prediction. We will use two complementary approaches. 1) First, directed acyclic graphs of the use cases will be developed and used to guide debiasing techniques to improve the validity of the use case clinical risk prediction tools. 2) Second, we will also identify cases where debiasing techniques are not sufficient and propose characteristics that can be used to identify these patients at the point of care to facilitate real-time decision support to reduce bias and improve care.

Hypothesis 1a: Risk prediction models that incorporate appropriate adjustment for data quality and SDoH concepts are *significant* predictors of actual outcomes.

Hypothesis 1b**:** Risk prediction models that incorporate appropriate adjustment for data quality and SDoH concepts are *better* predictors of actual outcomes than the original risk scores.

Hypothesis 2: Risk model failure can be predicted by factors that are easily assessed at the point of care

C5a. Structural debiasing model development and testing

Data for model development and initial testing will be limited to OCHIN clinics with an SDoH screening rate of at least 85% (**C2bii**).[100, 102] All analyses described here will be performed in SAS version 9.4.

Based on the results from **Aims 1** and **2**, we will generate separate directed acyclic graphs (DAGs) for each of the use cases.[52] DAGs are graphical nonparametric probabilistic diagrams that depict presumed causal relationships and can be used to identify "biasing pathways" that inhibit valid causal inference, to select variables for confounding control, and to adjust for selection bias and non-representativeness.[51-58] In Figure 4, bias occurs because the selection node, poor data quality, which is influenced by SDoH and utilization, and other variables in the DAG, is a "collider variable," and conditioning on this collider by selecting patients with sufficient data quality creates "collider stratification bias."[120] This type of bias can sometimes be mitigated by adjusting for confounder variables that also influence selection, but if the exposure and outcome also predict selection, confounder adjustment may not be sufficient to control all the bias. If it is possible to quantify the selection mechanism, the biasing pathways may be blocked by weighting the outcome model in a procedure called inverse probability of selection (IPS) weighting.[75, 121, 122]

We will use an established method, inverse probability weighting[123] for adjustment of bias due to collider stratification bias (a form of selection bias that occurs due to the preferential inclusion of patients with higher data quality as seen in Figure 1),[124, 125] with data fusion[126] techniques that, while underused in the analysis of EHRs, has been previously applied to multiple types of data, including EHRs, by Dr. Thompson.[72, 75] This

adjustment can be done using externally obtained bias adjustment parameters (which we will estimate in **Aim 2** using our gold standard population), fused with data on all selected aim 3 patients, to impute the corresponding selection probability for each patient under the assumed selection and data generating mechanisms, as depicted in each use case DAG. Bias attributed to selection on (collider stratification) on data quality can then be adjusted using inverse probability weighted fitting of any planned outcome regression model. This work is an extension of inverse probability of censoring weighting[127] in prospective cohort studies, but rather than reliance on data from a censored population, we will use data fusion[126, 128] to combine our validated, empirical estimates from **Aim 2** as selection bias adjustment parameters for clinical risk prediction in a separate population.[57] The IPS-weighted results will reflect a pseudopopulation that up/down-weights under/over-represented individuals based on important predictors of data quality. This method is flexible to a variety of collider bias scenarios and has been successfully applied to address systematic error and non-representativeness in EHRs.[75]

C5b. Validation and Error Analysis

The resulting models from C5a will be applied to a random sample of 10,000 OCHIN patients drawn from the full OCHIN patient population and basic metrics of predictive accuracy will be calculated. Actual clinical outcomes will be drawn from a combination of OCHIN clinical data, claims data, and death index data. If the models are able achieve good performance in this validation sample it will demonstrate the generalizability and transferability of our models and approach. While the majority of OCHIN patients have basic demographic concepts recorded, SDoH screening rates in the overall OCHIN network are approximately 10%, and not all SDoH domains are addressed in all screenings.[129] These data will more accurately reflect the current state of SDoH screening in non-OCHIN healthcare settings.

We will conduct an error analysis of cases for which the models do not predict correctly to identify shared characteristics (clinical or SDoH) of patients that are associated with a high chance of risk prediction failure. Common characteristics will be mapped to corresponding concepts within the EHR that are easily accessible to the point of care. Finally, we will determine the positive predictive value and sensitivity of applying this list of patient characteristics in identifying patients where risk prediction failure is likely. These characteristics will provide an opportunity, potentially through automated systems, for providers to identify these at-risk patients and inform appropriate clinical decision making.

9) Privacy, Confidentiality and Data Security

Because this will be a fully retrospective study relying on analysis of existing real world data, there will be no privacy or confidentiality concerns arising from recruitment, consent, or study procedures, but there will be some risks arising from secondary use of clinical data. We will seek a waiver of consent to use these data, and will mitigate these risks to the best of our ability.

All three sites will abide by relevant federal and institutional regulations to ensure patient privacy and confidentiality. We have attached our Protection of Human Subjects document, which provides extensive details about the steps each site will take.

In short, only study team members who have completed human subjects and conduct of research trainings will have access to the data. Data will be stored and transmitted using secure, institutionally-approved tools (e.g., OneDrive at OHSU, or OHSU-managed computers). Study team members will only be able to access the data using their institutional network login information (username and passwords).

We will not be able to fully deidentify the data, since we will require dates, but identifiers will be stripped from the data where possible.

10) Risks and Benefits

a) Risks to Subjects

The risks to subjects are primarily involving threats to privacy and confidentiality of personal health information. As described above, we will take all necessary and appropriate precautions to mitigate this risk.

b) Potential Benefits to Subjects

We do not anticipate any direct benefit to patients whose data will be used.

OREGON HEALTH & SCIENCE UNIVERSITY

IRB MEMO

Research Integrity Office
3181 SW Sam Jackson Park Road - L106RI
Portland, OR 97239-3098
(503)494-7887 irb@ohsu.edu

APPROVAL OF SUBMISSION

March 7, 2024

Dear Investigator:

On 3/7/2024, the IRB reviewed the following submission:

IRB ID:	STUDY00025986	MOD ID:	MOD00055093
Type of Review:	Modification / Update		
Title of Study:	Health equity and the impacts of EHR data bias associated with social determinants		
Title of modification	Revised protocol with OHA APAC info		
Principal Investigator:	Nicole Weiskopf		
Funding:	Name: DHHS NIH Natl Library of Medicine, PPQ #: 1022094		
IND, IDE, or HDE:	None		
Documents Reviewed:	• Protocol		

The IRB granted final approval on 3/7/2024. The study requires you to submit a check-in before 7/10/2026.

Review Category: Expedited-Minor Modification

Copies of all approved documents are available in the study's **Final** Documents (far right column under the documents tab) list in the eIRB. Any additional documents that require an IRB signature (e.g. IIAs and IAAs) will be posted when signed. If this applies to your study, you will receive a notification when these additional signed documents are available.

Ongoing IRB submission requirements:

- Six to ten weeks before the eIRB system expiration date, submit a check-in.
- Any changes to the project must be submitted for IRB approval prior to implementation.
- Reportable New Information must be submitted per OHSU policy.
- Submit a check-in to close the study when your research is completed.

Guidelines for Study Conduct

In conducting this study, you are required to follow the guidelines in the document entitled, "Roles and Responsibilities in the Conduct of Research and Administration of Sponsored Projects," as well as all other applicable OHSU IRB Policies and Procedures.

Requirements under HIPAA

If your study involves the collection, use, or disclosure of Protected Health Information (PHI), you must comply with all applicable requirements under HIPAA. See the HIPAA and Research website and the Information Privacy and Security website for more information.

IRB Compliance

The OHSU IRB (FWA00000161; IRB00000471) complies with 45 CFR Part 46, 21 CFR Parts 50 and 56, and other federal and Oregon laws and regulations, as applicable, as well as ICH-GCP codes 3.1-3.4, which outline Responsibilities, Composition, Functions, and Operations, Procedures, and Records of the IRB.

Sincerely,

The OHSU IRB Office

Please answer each of the following questions about APAC data request options:

Please indicate the year(s) of data requested	2011-2022 X
Do you want out-of-state people and their claims included?	No X
Do you want orphan claims included? (claims, but no eligibility or coverage reported)	Yes X
Do you want coordination of benefit claims included?	Yes X
Do you want self-insured commercial data included?	Yes X
Do you want PEBB and OEBB commercial data included?	Yes X
What payer types do you want?	Commercial, Medicaid, Medicare Adv X
Do you want all medical claims?	Yes, all medical claims X

Please answer each of the following questions about APAC data request options:	
How do you want medical claim type(s) identified and selected?	APAC definition (see data element: APAC grouper) **X**
Which medical claim types do you want?	All claims **X**
Do you want pharmacy claims?	Yes **X**
Do you want dental claims?	No **X**
Do you want monthly eligibility data (insured/covered months by plan)?	Yes **X**
Do you want member demographic data?	Yes and I did not request monthly eligibility data **X**
Do you want provider data?	No **X**
Do you want claims and eligibility data for selected age groups only?	Specify age range: **Age 18 and older as of Dec 31, 2022**
Do you want to limit claims and eligibility data by gender?	Include all **X**

Please answer each of the following questions about APAC data request options:	
Do you want to limit medical claims data to selected diagnoses?	No **X**
Do you want to limit pharmacy data to selected NDC codes or therapeutic classes?	Yes. Please list NDC or therapeutic class codes **See Attached Sheet with Medications**
Are you requesting identifiable data?	No **X**
One payer reported the claim status for all of their claims as fee-for-service for some years when most claims were encounter or managed care claims. Do you want the claim status changed to managed care?	Change to encounter **X**
Do you want APAC data linked to Oregon Center for Health Statistics (CHS) Death Certificate data? You will need approval from both CHS and APAC. Submit request to APAC first and after approval submit request to CHS and provide APAC approval notice	No **X**
Is your requested APAC data going to be linked by APAC Team or data requester with any other data source?	Yes **Requesting HSRI/NORC Linkage to OCHIN EHR**

Field Requested	Data Element	Security Level	Description	Justification/use within specific project
The data elements highlighted in blue are provided in every data request	uid	De-Identified	A unique identifier that links to the row as submitted in the MC Intake File Layout. Used for linking tables/views	Not required
	release_id	De-Identified	A value associated with the data release	
	mc059_service_start_dt	De-Identified	Date services for patient started	
	dw_claim_id	De-Identified	A unique medical claim identifier	
	mc005_line_no	De-Identified	Line number for the claim that begins with 1 and is incremented by 1 for each additional service line of a claim	
	uniquepersonID	De-Identified	A unique identifier for a person across payers and time	
	dw_member_id	De-Identified	A payer & plan specific unique identifier for a person. A person can have multiple member IDs for a single payer because they can have multiple plans. DW_member_IDs are not unique identifiers for a person across payers and years	
	dw_person_id	De-Identified	Vendor identifier for a person across payers and time-many people have more than one assigned identifier	
	mc038a_cob_status	De-Identified	Coordination of benefit claim. Indicates secondary payer for a claim	
	orphan_fl	De-Identified	Identifies orphan claim with no corresponding eligibility for the date of service	
	mc003_insurance_product_type_cd	De-Identified	A code that indicates an insurance coverage type. Data element required for linking claims to member months	

	me016_member_state	De-Identified	Member State from latest quarterly data submitted	
X	mc060_service_end_dt	De-Identified	Date services for patient ended	We will use care start and stop dates to identify periods of readmission as a main study outcome
X	Claim_LOB	De-Identified	Payer line of business: 1 (Medicare), 2 (Medicaid), 3 (Commercial, 0 (no line of business reported), -99 (duplicate data reported)	We will treat insurance type as a covariate in predicting our outcomes.
	self_insured_fl	De-Identified	Self Insured flag	
X	mc018_admit_dt	De-Identified	Admission date	We will use care start and stop dates to identify periods of readmission as a main study outcome
X	mc203_admit_type_cd	De-Identified	Admission type:1 (Emergency), 2 (Urgent), 3 (Elective), 4 (Newborn), 5 (Trauma Center), 9 (missing)	We will need to discern which admissions were for elective vs. emergency procedures
X	mc205_admit_diagnosis_cd	De-Identified	Admitting diagnosis. ICD-10 diagnosis code for dates of service beginning 10/01/2015, ICD-9 diagnosis code for dates of service before 10/01/2015	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc070_discharge_dt	De-Identified	Discharge date-required for inpatient hospitalization	We will use care start and stop dates to identify periods of readmission as a main study outcome

X	mc023_discharge_status_cd	De-Identified	Status for member discharged from a hospital	Mortality is a study outcome; we will need to know if a patient was alive at discharge
X	LOS	De-Identified	Length of stay of inpatient admission measured in days. Discharge Date - Admit Date. <1 is rounded to 1. Negative values set to NULL	We will need length of stay information to differentiate hospital stays and identify readmissions
X	mc036_bill_type_cd	De-Identified	Type of bill on uniform billing form (UB)	We will need UB Codes to identify emergency visits vs. elective visits
X	mc037_place_of_service_cd	De-Identified	Industry standard place of service code	We will need Place of Service to differentiate emergency and elective visits
X	mc054_revenue_cd	De-Identified	Revenue code	We will need Revenue Codes to identify emergency visits vs. elective visits
X	mc041_principal_diagnosis_cd	De-Identified	Principal Diagnosis code	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	Dx_Description	De-Identified	ICD diagnosis code description	We will need this to discern admissions for cardiovascular disease and events and related conditions

X	Dx_Type	De-Identified	ICD diagnosis code type	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc041p_poa_p	De-Identified	Required present on admission flag for diagnosis 1: Yes, no, W (clinically undetermined), U (information not in record), diagnosis exempt from POA reporting (1), Null if not reported	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	POA_Description	De-Identified	Present on admission description	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc042_other_diagnosis_2	De-Identified	Additional Diagnosis 2	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc042p_poa_2	De-Identified	Required POA flag for diagnosis 2 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment

X	mc043_other_diagnosis_3	De-Identified	Additional Diagnosis 3	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc043p_poa_3	De-Identified	Required POA flag for diagnosis 3 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc044_other_diagnosis_4	De-Identified	Additional Diagnosis 4	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc044p_poa_4	De-Identified	Required POA flag for diagnosis 4 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc045_other_diagnosis_5	De-Identified	Additional Diagnosis 5	We will need this to discern admissions for cardiovascular disease and events and related conditions

X	mc045p_poa_5	De-Identified	Required POA flag for diagnosis 5 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc046_other_diagnosis_6	De-Identified	Additional Diagnosis 6	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc046p_poa_6	De-Identified	Required POA flag for diagnosis 6 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc047_other_diagnosis_7	De-Identified	Additional Diagnosis 7	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc047p_poa_7	De-Identified	Required POA flag for diagnosis 7 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment

X	mc048_other_diagnosis_8	De-Identified	Additional Diagnosis 8	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc048p_poa_8	De-Identified	Required POA flag for diagnosis 8 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc049_other_diagnosis_9	De-Identified	Additional Diagnosis 9	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc049p_poa_9	De-Identified	Required POA flag for diagnosis 9 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc050_other_diagnosis_10	De-Identified	Additional Diagnosis 10	We will need this to discern admissions for cardiovascular disease and events and related conditions

X	mc050p_poa_10	De-Identified	Required POA flag for diagnosis 10 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc051_other_diagnosis_11	De-Identified	Additional Diagnosis 11	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc051p_poa_11	De-Identified	Required POA flag for diagnosis 11 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc052_other_diagnosis_12	De-Identified	Additional Diagnosis 12	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc052p_poa_12	De-Identified	Required POA flag for diagnosis 12 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment

X	mc053_other_diagnosis_13	De-Identified	Additional Diagnosis 13	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc053p_poa_13	De-Identified	Required POA flag for diagnosis 13 if populated	We will need this to differentiate preexisting conditions from those that developed or worsened during treatment
X	mc201_icd_version_cd	De-Identified	Identifies ICD9 or ICD10 version	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mc055_procedure_cd	De-Identified	Current Procedural Terminology (CPT) code or Healthcare Common Procedure Coding System (HCPCS)	We will need this to identify cardiovascular disease and events
X	Px_Type	De-Identified	ICD procedure code type	We will need this to identify cardiovascular disease and events
X	CPT description	De-Identified	Short Description of Current Procedural Terminology, created and owned by the American Medical Association	We will need this to identify cardiovascular disease and events
X	consumer_friendly_descriptor	De-Identified	Consumer Friedly description of Current Procedural Terminology, created and owned by the American Medical Association	We will need this to identify cardiovascular disease and events
X	mc056_procedure_modifier_1_cd	De-Identified	CPT or HCPCS modifier	We will need this to identify cardiovascular disease and events

X	mc057_procedure_modifier_2_cd	De-Identified	CPT or HCPCS modifier	We will need this to identify cardiovascular disease and events
X	mc057a_procedure_modifier_3_cd	De-Identified	CPT or HCPCS modifier	We will need this to identify cardiovascular disease and events
X	mc057b_procedure_modifier_4_cd	De-Identified	CPT or HCPCS modifier	We will need this to identify cardiovascular disease and events
X	modifier description	De-Identified	Description of Outpatient Procedure modifier code, from either CPT, HCPC, or Ambulance code list.	We will need this to identify cardiovascular disease and events
X	APACgrouper	De-Identified	Groups all lines of a claim in prioritized order as inpatient, emergency department, outpatient, professional, pharmacy and other based on type of bill, revenue and place of service codes	We will need this to differentiate cardiovascular events, treatments, and readmissions
X	claim_type	De-Identified	Vendor generated claim ltype. Identifies claim lines as inpatient facility claim (1), outpatient facility claim (2) and professional claim (3) based on bill type, revenue code and place of service. Null means claim line type could not be determined.	Please only include professional claims
X	mc058_icd_primary_procedure_cd	De-Identified	The main inpatient procedure code	We will need this to differentiate cardiovascular disease and events
X	mc058a_icd_procedure_2	De-Identified	Inpatient procedure ICD-10 code 2	We will need this to differentiate cardiovascular disease and events

X	mc058b_icd_procedure_3	De-Identified	Inpatient procedure ICD-10 code 3	We will need this to differentiate cardiovascular disease and events
X	mc058c_icd_procedure_4	De-Identified	Inpatient procedure ICD-10 code 4	We will need this to differentiate cardiovascular disease and events
X	mc058d_icd_procedure_5	De-Identified	Inpatient procedure ICD-10 code 5	We will need this to differentiate cardiovascular disease and events
X	mc058e_icd_procedure_6	De-Identified	Inpatient procedure ICD-10 code 6	We will need this to differentiate cardiovascular disease and events
X	mc058f_icd_procedure_7	De-Identified	Inpatient procedure ICD-10 code 7	We will need this to differentiate cardiovascular disease and events
X	mc058g_icd_procedure_8	De-Identified	Inpatient procedure ICD-10 code 8	We will need this to differentiate cardiovascular disease and events
X	mc058h_icd_procedure_9	De-Identified	Inpatient procedure ICD-10 code 9	We will need this to differentiate cardiovascular disease and events
X	mc058j_icd_procedure_10	De-Identified	Inpatient procedure ICD-10 code 10	We will need this to differentiate cardiovascular disease and events

X	mc058k_icd_procedure_11	De-Identified	Inpatient procedure ICD-10 code 11	We will need this to differentiate cardiovascular disease and events
X	mc058l_icd_procedure_12	De-Identified	Inpatient procedure ICD-10 code 12	We will need this to differentiate cardiovascular disease and events
X	mc058m_icd_procedure_13	De-Identified	Inpatient procedure ICD-10 code 13	We will need this to differentiate cardiovascular disease and events
X	mc201_icd_version_cd	De-Identified	ICD version code 9 - ICD-9, 10 - ICD-10	We will need this to differentiate cardiovascular disease and events
X	final_mdc	De-Identified	a code identifying the final Major Diagnostic Category (MDC)	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	final_drg	De-Identified	a code indentifying the final Diagnosis Related Group	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	final_ms_ind	De-Identified	a flag indicating if final_mdc is medical or surgical	We will need this to discern admissions for cardiovascular disease and events and related conditions

X	drg description	De-Identified	Final DRG description	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	mdc description	De-Identified	Final MDC description	We will need this to discern admissions for cardiovascular disease and events and related conditions
X	MS DRG MDC cross walk Description	De-Identified	Crosswalk DRG to MDC	We will need this to discern admissions for cardiovascular disease and events and related conditions

Field Requested	Data Element	Security Level	Description	Justification/use within specific project
The data elements highlighted in blue are provided in every data request	uid	De-Identified	A unique identifier that links to the row as submitted in the PC Intake File Layout. Used for linking tables/views	Not required
	release_id	De-Identified	A value associated with the data release	
	dw_claim_id	De-Identified	A unique medical claim identifier	
	pc032_prescription_fill_dt	De-Identified	Prescription fill date	
	dw_member_id	De-Identified	A payer & plan specific unique identifier for a person. A person can have multiple member IDs for a single payer because they can have multiple plans. DW_member_IDs are not unique identifiers for a person across payers and years	
	uniquepersonID	De-Identified	A unique identifier for a person across payers and time	
	dw_person_id	De-Identified	Vendor identifier for a person across payers and time-many people have more than one assigned identifier	
	me016_member_state	De-Identified	Member State from latest quarterly data	
	orphan_fl	De-Identified	Identifies orphan claim with no corresponding eligibility for the date of service	
X	pc003_insurance_product_type_cd	De-Identified	A code that indicates an insurance coverage type	We will treat insurance type as a covariate in predicting our outcomes.

X	pc001_payer_type	De-Identified	Payer reported payer type codes:(C) Carrier, (D) Medicaid, (G) Other government agency, (P) Pharmacy benefits manager, (T) Third-party administrator, (U) Unlicensed entity	We will treat insurance type as a covariate in predicting our outcomes.
X	Claim_LOB	De-Identified	Payer line of business: 1 (Medicare), 2 (Medicaid), 3 (Commercial, 0 (no line of business reported), -99 (duplicate data reported)	We will treat insurance type as a covariate in predicting our outcomes.
X	pc026_drug_cd	De-Identified	National Drug Code (NDC)	We will need this to differentiate the prescription drug
X	pc033_dispensed_qty	De-Identified	Quantity dispensed	We will need this to differentiate the prescription drug
X	pc028a_alt_refill_no	De-Identified	Alternate refill number	We will need this to estimate the treatment duration with the prescription drug
X	pc034_days_supply_qty	De-Identified	Number of days that the drug will last if taken at the prescribed dose	We will need this to estimate the treatment duration with the prescription drug
X	pc028_calc_refill_no	De-Identified	Processor's count of times prescription refilled	We will need this to estimate the treatment duration with the prescription drug

New or Amended APAC Data Request Review (custom or OHA Business Associate)

Staff Reviewer: Oliver

DRTS Number: 6250

Date review completed: 5/10/2024

	Yes	No	N/A	Need more information
Is this a new APAC request?	X			
New APAC Request (skip to next section if amendment request):				
1.1 Project staff contact information provided	X			
1.2 Project technical staff information provided		X		
2.1 Project summary provided with adequate detail to identify a specific unambiguous project	X			Investigate whether adverse SDoH are drivers of poor EHR data quality.
2.2 Research questions provided with adequate detail	X			Three specific aims in APAC-3 and research protocol.
2.3 Described planned products and reports derived from requested data	X			Key product is improved risk prediction algorithms for six widely adopted clinical risk prediction tools.
2.4 Project begin and end date provided	X			End 4/30/2027
2.5 Acknowledgement that APAC data cannot be reused beyond the DUA	X			
2.5 Acknowledgement that data cannot be shared beyond the DUA	X			
3.1ab Data request purpose box checked & description	X			
3.2 Checked box for level of data identifiers	X			
3.3 IRB application, approval memo, end date	X			Approved 3/7/2024,End 7/10/2026
4.1 Completed data elements workbook	X			
4.2 Adequately described how the data elements requested are the minimum necessary	X			
5.1 Plan provided to prevent re-identification	X			
5.2ab Plan to link APAC data to other data source	X			Member/Client linkage to OCHIN EHR and SDoH data
5.2c Requests OHA to link APAC to other data	X			OCHIN will send finder file to HSRI
5.2d Detailed data linking plan provided	X			Will use standard HSRI process. Output is crosswalk from APAC dw_member_id to OCHIN Client_ID.
5.3 Provided adequate description of data management, security and data destruction plan	X			

Passes Minimum Necessary Review	**X**			Limited to age 18+, linked to OCHIN, and (specific diagnoses OR specific NDCs OR inpatient/ED admit in CYs 2019-2022). Narrow medical claims request with no payment or provider fields. Requested only 8 pharmacy fields.
Recommend management approval	**X**			