Consortium

The **American Family Cohort (AFC) Research Consortium**, which was established in 2019, arose from a collaboration between the **American Board of Family Medicine (ABFM)** and the **Stanford Center for Population Health Sciences (PHS)**. Recognizing the potential of the data, and the alignment of mission between ABFM and the PHS, the organizations partnered to make the ABFM PRIME data available for primary care and population health research. PHS had made considerable investments in data management strategies and secure computational infrastructure which optimized for secure multi-institutional data sharing. PHS became the custodian of the research version of ABFM PRIME – the **American Family Cohort (AFC)**. The consortium has expanded to include new partners in the core data management function (Census Bureau) and research teams from multiple institutions (CDC, other academic research teams and organizations).

American Family Cohort (AFC) data

The PRIME registry is the source of the AFC data. PRIME is sponsored by the American Board of Family Medicine (ABFM) whose objective was to establish a Qualified Clinical Data Repository (QCDR) for primary care. The registry provides tools to evaluate primary care practice performance, support population health and risk stratification, improve primary care practice as well as patient outcomes, and alleviate Centers for Medicare and Medicaid Services (CMS) reporting for their payment programs. The PRIME registry, certified by CMS in 2016, represents over 3,000 active clinicians representing 50 states from data on over eight million patients. Dating back to 2010, the ABFM PRIME registry is the largest clinical registry for primary care in the nation*.*

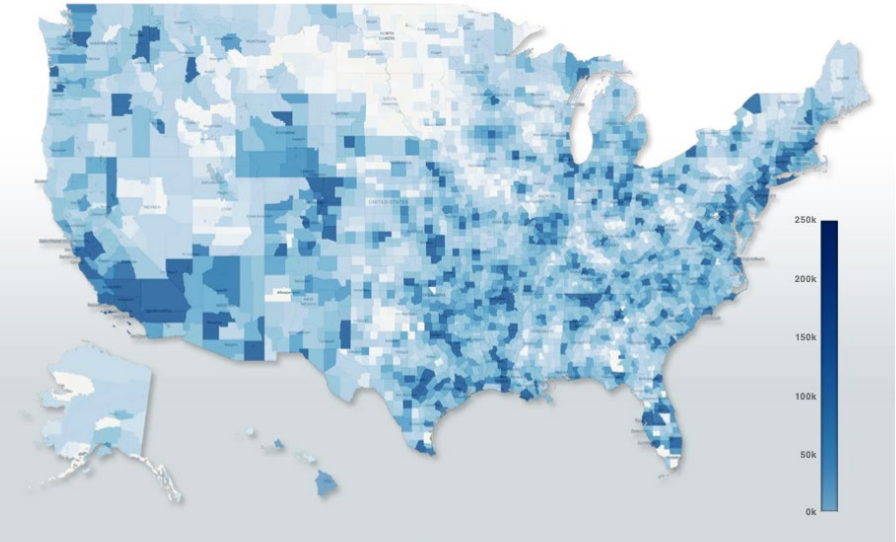
Creating the American Family Cohort (AFC) from the PRIME registry data. Since 2019, the American Board of Family Medicine has partnered with the Stanford University Center for Population Health Sciences to create the American Family Cohort from the PRIME registry data. The PRIME registry consists of data that is continuously collected from the PRIME Registry and transformed into research data.

Data are electronically extracted directly from the electronic health records (EHRs) via online portals. Data elements include both structured and unstructured data, typical of disparate EHRs. Data elements include patient demographics, diagnoses and interventions for the patients such as medications and therapies, encounter-specific data, patient-reported outcomes (PROs), and some limited clinician-specific details. All data are collected during routine assessment of clinical care of patients whose main goals are to support practice-specific quality improvement activities as well as CMS-specific quality reporting for payment. The PRIME registry includes National Qualify Forum (NQF)-endorsed measures and a patient-reported outcome (PRO) measure tool that aids in tracking practice performance.

Storage, security, and computation on such EHR data are challenging, to the extent that the data can be unworkable and impractical for the types of statistical models that are required for the standards expected within medicine and for regulatory purposes. Our data team (led by Ayin Vala) employs individuals with specific skills in handling and researching this type of large and complex EHR data. The EHR data extracted is obtained in the form of hundreds of thousands of raw parquet files, a common big data file format. Data is then normalized by our team into predefined tables to ensure uniformity across data from thousands of facilities. Collecting data this way is more cumbersome, but it is used to ensure data integrity and prevent mistakes from happening from upstream data aggregation. Securing such sensitive EHR data is a top priority. Data are protected by extensive procedural, regulatory and technical controls. Data are stored encrypted at rest by default, using multiple layers of encryption. The dataset consists of several billion records. Consequently, in order to perform the necessary data curation and computations necessary for the research, a highly scalable data warehouse system is required. Stanford uses a secure cloud environment to manage the large volume of data, powered by thousands of cores, petabytes in storage capacity, and terabytes in networking bandwidth. Several enhancements to the AFC data are either underway or will be initiated this year including transformation to the OMOP common data model and linkage to Medicare and Medicaid claims data. The AFC data include:

Good representation from populations that have been disproportionately impacted by SARS-CoV-2 infection from all 50 states (Figure 1), with enrollment roughly proportional to state populations. AFC includes patients on private insurance plans, Medicaid and Medicare, increasing the representation of vulnerable populations, and the generalizability of the sample to the overall US population.

Figure 1: Number of AFC patients by county



Racially and ethnically diverse data include 540,000 Black patients, 150,000 Asian patients, 51,000 Native American and Alaska Native patients and 16,000 Native Hawaiian and Pacific Islander patients. The remaining 4.8 million patients are White, and 758,000 patients have identified as Hispanic or Latino. The diversity is a major strength for addressing all our key questions which focus on underserved and marginalized populations. We also are able to disaggregate racial and ethnic identifiers into more specific categories because this data is collected as free text. While not part of this current proposal, future work will for example be able to examine Asian and Latino subgroups for health inequities.

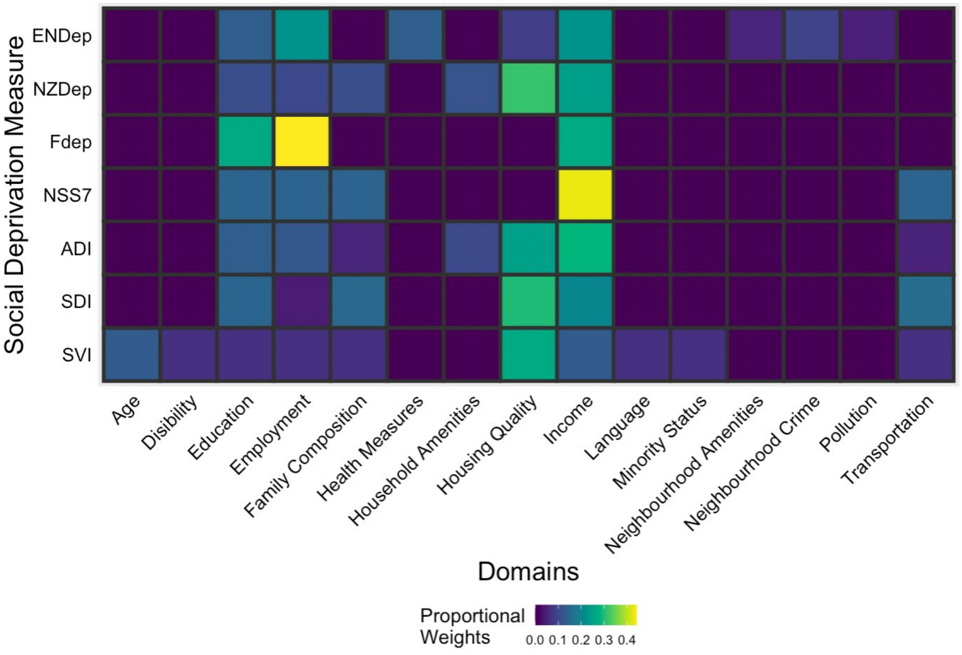
Approximately 900,000 children whom we are able to link to parent health records, including a large number from minority populations from practices in *rural areas* of the US, allowing for analyses across the life course, and across generations in diverse groups.

These data have been used to study COVID-19 with robust sample sizes and ability to detect differences by race, ethnicity, and neighborhood social deprivation, as we describe below in the preliminary data section.[17]

Preliminary data

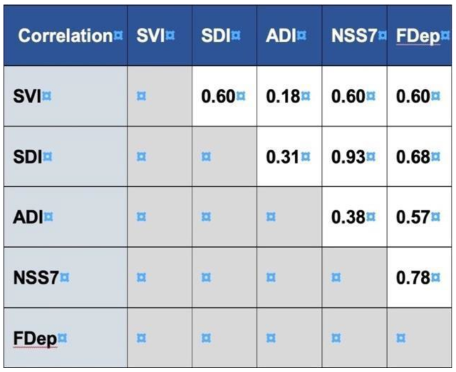
Preparation of Social Deprivation Indexes Dr. Rehkopf and the AFC Consortium have prepared six social deprivation indexes including the three that will be used for this project and made the available with accompanying crosswalk files for overlay by geographic unit (census tract, zip code or PUMA) with health or other outcome data.[18]

Figure 2: Comparison of factor weights for 6 social deprivation measures



Although there is considerable overlap between the factors used in these measures, there are important differences both in which elements are used and how they are weighted. Figure 2 shows the relative importance of 15 different factors used across six commonly used deprivation measures. The lighter colors show housing quality and income are a part of many of the measures, but also the extent to which the weighting across measures vary. Figure 3 shows that although many of the domains are shared across measures, the correlation of these deprivation indexes is relatively modest. For example, the ADI and SDI are only correlated at 0.31, thus it is highly plausible that choice of measure will impact analysis estimates, something that we will evaluate in this project. An outcome of our proposed work here will be to compare the use of the three most commonly used measures for understanding the source of health inequities in the AFC data.

Figure 3: Correlation (r^2) between predictive power of each of 6 social



Correlation	SVI	SDI	ADI	NSS7	FDep
SVI		0.60	0.18	0.60	0.60
SDI			0.31	0.93	0.68
ADI				0.38	0.57
NSS7					0.78
FDep					

Data Elements in AFC

The AFC data include key patient and primary care attributes of interest. These can be divided into five major categories: (1) health-need attributes, (2) race/ethnicity, (3) social deprivation indexes, (4) regional and county-level resource indicators, and (5) clinical presentation and treatment.

Health-need Attributes Patient attributes include patient's sex at birth, age of the patient that can be derived from birth date, diagnoses of chronic conditions via encounter-specific diagnosis codes and history of visits. Diagnoses of chronic conditions would be obtained via *International Classification of Diseases, 10th Revision* diagnosis codes in any position during the primary care visit.

Table 4: Race/ethnicity distribution in AFC

Race/ethnicity	Number of patients	Percent*
Non-Hispanic White	4,588,980	75.2
Non-Hispanic Black	540,023	8.8
Hispanic	757,990	12.4
Asian	149,211	2.4
American Indian/Alaska Native	51,130	0.8
Native Hawaiian/Pacific Islander	16,632	0.3
Multiple Races	2,113	0.03
Other race/not reported	1,660,672	

* among those with reported race/ethnicity

Race/Ethnicity Race/ethnicity is available on more than 80% of patients in AFC (Table 4). Race/ethnicity classifications include White, Asian, Hispanic or Latino, Black or African American, Native American or Alaska Native, Native Hawaiian or Other Pacific Islander, and Unknown/Missing/Not Available. Ethnicity is defined as Hispanic or Latino and Not Hispanic or Latino. Given the frequency distribution of patient's race/ethnicity, conditional on sample size in each race group, measure of differences in safety and efficacy of treatment may be relegated to those with the greatest representation (e.g. White, Black or African American, Hispanic or Latino). A primary analysis will include the research team recoded raw race/ethnicity variable available in AFC and a subsequent sensitivity analysis will be conducted where imputation methods applied to the race/ethnicity variable.

PHS also does ongoing work to curate and clean the race and ethnicity data. This will have substantial advantages for reclassifying "other race" as "multiple race" in order to more accurately classify persons who are multiracial.

Race/Ethnicity Imputation Excluding patients whose race/ethnicity is Unknown/Missing/Not Available from case analyses has been shown to underestimate the true levels of a disparity of interest.[20] To address this possibility, we impute race/ethnicity using Bayesian Improved First Name, Surname, and Geocoding (BIFSG),[21] which combines known statistics about name, location, and race (based on census and administrative records) with conditional independence assumptions to assign subjects a probability of being each of five race/ethnicity groups as defined by the original 1977 U.S. Office of Management and Budget (OMB) Standards for the Classification of Federal Data on Race and Ethnicity: Asian American and Pacific Islander (API), American Indian and Alaskan Native (AIAN), Black, Hispanic, White, and Other. We geocode AFC address information to the census block group (CBG) level using a combination of the Census Geocoder and Google Geocoding API. For prior distributions of race by CBG, we use 2018 5-yr summary data from the American Community Survey. For prior distributions of race by surname, we use the U.S. Census Bureau 2010 surname tables. For prior distributions of race by first name, we use the dataset published by Tzioumis, Konstantinos, 2018.[22] The only individuals we don't impute for are those missing a last name and state.

Social deprivation indexes The PRIME Registry includes geographic granularity that can afford linkages to external data sources for including social deprivation indexes as well as regional resource indicators. The Social Vulnerability Index (US Centers for Disease Control and Prevention) SVI was originally developed to assess preparedness for disasters[28,29] and has subsequently in prediction of health behaviors [30] and more recently in COVID-19 research.[31–35] These indexes are linked on the PHS data portal and described briefly below.

The CDC SVI is derived from U.S. Census data at the census tract level. The CDC uses 16 factors such as percent poverty, access to a vehicle, housing quality (crowding) and several others. These factors are then grouped into four themes which include socioeconomic status, household characteristics (household composition), racial and ethnic minority status and housing and transportation type.[36]

The Area Deprivation Index (the Robert Graham Center) ADI was originally constructed to model health service and health outcomes use and to better understand the impact of geography on the stability of models.[7]. Although the RGC ADI was modeled after indices from nations such as the United Kingdom and New Zealand which have used social deprivation indexes as a tool in allocation of public health resources for decades, the factors used and the weights assigned to factors are significantly different.[8] These indexes are used to identify areas at risk for high health care utilization and poor outcomes “hot spots” and communities limited resources and health services needs which are unmet “cold spots”.[37]

The Multidimensional Deprivation Index (U.S. Census Bureau) MDI is derived from Census data between 2010 and 2019 and include monetary and non-monetary factors. The measure is used include: standard of living, education, health, economic security, housing, and neighborhood. The Census has produced both national and state level indexes at the PUMA level. PUMAs are statistical geographic units used to ensure that all areas have at least 100,000 people. The boundaries of PUMAs do not necessarily coincide with counties. The MDI is used primarily to estimate poverty and for other applications in social sciences. Census is interested in improving and refining the MDI.[38,39]

The Social Deprivation Index (SDI), available from the American Community Survey is a composite measure of area-level deprivation on seven demographic characteristics that can be derived from the Zip Code Tabulation Area (ZCTA) or Primary Care Service Area (PSCA). These 7 measures include percent single-parent household, percent living in the rented housing unit, percent living in the overcrowded housing unit, percent of households without a car, and percent non-employed adults under 65 years of age. SDI will be considered based on quintiles or quartiles.[7,8,40] Insurance type at baseline is also included, defined in unstructured, free-text data, but may be considered. Insurance information in PRIME includes the name of the plan, insurance company name, plan type (e.g. Health Maintenance Organization (HMO), Preferred Provider Organization (PPO), etc.) and other information salient for billing purposes.[41] We will consider the use of the National Neighborhood Data Archive (NaNDA) data elements on affluence, including proportion of adults with a college education, incomes >$75K, and people with managerial and occupational professions.[40,42]

Regional and County-Level Resource Indicators Using the patient U.S. postal zip code available in the PRIME, we will implement rural-urban commuting area (RUCA) codes available from the U.S. Department of Agriculture to define patient rurality.[43] In addition, relevant county-level health care resource indicators from the Area Resource File (ARF) include number of physicians per 1,000 people and number of hospitals per 100,000 people as has been previously described and used in other work focused on vulnerable populations or those with disadvantages.[44–46] To assign these regional resource measures, we will use the Federal Information Processing System (FIPS) county codes to assign to patients in AFC.