The Intersection of Identity and Data

A Statistical Approach to Disaggregate Ethnic Identities in Colorado

Methods, Lessons Learned, and next Steps

January 2023

COLORADO HEALTH INSTITUTE

Informing Strategy. Advancing Health.

Executive Summary

More than one in five Coloradans (21%) identify as Hispanic or Latino, based on the 2021 Colorado Health Access Survey (CHAS). However, data on more specific ethnic identities within this group, such as Chicano or Central American, are limited. The lack of available data diminishes the ability to understand how health outcomes, access to care, and use of care may differ among specific Hispanic or Latino ethnic groups in Colorado.

Data disaggregation is a set of methods used to uncover populations often hidden in the data. The term describes the process of collecting and analyzing information on granular subcategories of people — often grouped by race and ethnic identity — that can reveal disparities where aggregated data cannot.

To bridge gaps in data reporting, the Colorado Health Institute (CHI), in partnership with the University of California, Los Angeles (UCLA) Center for Health Policy Research, and the Robert Wood Johnson Foundation, explored a strategy of retrospectively disaggregating Hispanic or Latino ethnicity across three Colorado health datasets: a survey, insurance claims, and electronic health records.

CHI approached this work in two phases. Phase I assessed the feasibility of methods that CHI and other entities could use to disaggregate data. In Phase II, CHI applied those methods across Colorado's existing data systems. Those systems are:

- Colorado Health Access Survey (CHAS)
- Colorado All Payer Claims Database (CO APCD)
- Colorado Health Observation Regional Database Service (CHORDS)

The 2021 CHAS included eight subidentity options for Hispanic/Latino participants, whereas CO APCD and CHORDS only had a Hispanic/Latino option for patients. Using predictive modelling, CHI assessed the relationship between key demographic characteristics collected on the 2021 CHAS and subidentity options within the Hispanic or Latino community. The question was whether demographic characteristics common to all three datasets could help predict Hispanic or Latino subidentities for patients in CO APCD and CHORDS. After validation of the regression models, results showed that models built for three of the subidentities in the CHAS — Caribbean/Central American, South American, and Spanish American — could be applied to the CHORDS and CO APCD datasets.

CHI reached out to Colorado community groups throughout the Phase II process to better understand their needs and wants related to disaggregated data. CHI thanks those community members for their participation. This engagement shaped the work, which was made better for it.

Key findings and reflections from the second phase of our analysis include:

- Data disaggregation methods should be community informed to ensure that the collection, management, and analysis of racial/ethnic identity data are correctly interpreted, actionable, and useful to community organizations' work and missions.
- Local data expertise should be leveraged not only to increase efficiency of the development of the statistical methods but also to increase awareness and emphasize the importance of disaggregating data within more data systems.
- Data collection efforts set the stage for this work — invest the time and resources into available data systems to gather representative disaggregated data.

In Phase III, CHI will apply the models to the CHORDS and CO APCD datasets to understand more about behavioral health in Colorado's specific communities. We plan to analyze behavioral health utilization patterns among the disaggregated Hispanic or Latino ethnic groups identified in Phase II.

Acknowledgements

Colorado Health Institute Team
Lindsey Whittington, Lead Author
Kristi Arellano
Jeff Bontrager
Brian Clark
Cliff Foster
Deanna Geldens
Emily Johnson
Paul Presken
Guadalupe Solís

Collaborators
Emily Bacon, PhD, Bacon Analytics LLC
Anna Furniss, MS, University of Colorado Anschutz Medical Campus
Maria de Jesus Diaz-Perez, PhD, Center for Improving Value in Health Care

The Colorado Health Institute thanks AJ Scheitler, EdD, Ninez Ponce, MPP, PhD, and Susan Babey, PhD, at the UCLA Center for Health Policy Research for their guidance and support of this project. CHI completed this assessment with support from the Robert Wood Johnson Foundation.

Table of Contents

Introduction

Members of racial and ethnic minority groups face historical and systemically rooted inequities in the United States that continue to propagate significant health disparities. In general, data sources used to analyze these disparities rely on five aggregated racial and ethnic categories: American Indian or Alaska Native, African American or Black, Asian American or Native Hawaiian or Other Pacific Islander, Hispanic or Latino/a, and White. Some data sources use more categories and others fewer. Because these categories are so broad, differences within these groups tend to be disregarded. In response, researchers have developed and used new approaches to disaggregate race and ethnicity data.

Data disaggregation is the method of separating larger groups into smaller populations to understand trends or patterns that might have otherwise gone undescribed. That way, health services researchers, for example, can use information about these subgroups to better understand differences in health outcomes, use of services, health behaviors, or barriers to accessing health care.

Concepts like culture, identity, race, national origin, and ethnicity are incredibly complex, are intersectional, and influence how someone interacts within our social environments. CHI has used this project to learn more about the Colorado context of these concepts and how they interact to understand more about how the data that we have currently collected are limited to a certain lens of these concepts. For this report, we have used the concept of culture to frame someone's ethnic/racial identity or language and how these aspects interact with other demographic characteristics, like gender. Ethnicity in this paper is specific to the Hispanic or Latino ethnic identity. Keeping all this in mind, CHI used these concepts to understand more about the communities that exist in our state.

Over the past two years, CHI, in partnership with the UCLA Center for Health Policy Research, has investigated the methods and feasibility of using data disaggregation techniques to understand more about Colorado's communities and the unique context that creates barriers to accessing care, poor health outcomes, and utilization of preventive and other health care services. Secondary objectives are to uplift the communities that are represented in the data and to inform the organizations, policymakers, and other decision-makers who might not be familiar with the issues facing these communities.

CHI used information captured on the 2021 CHAS to disaggregate data across two other data systems: the CO APCD and the CHORDS. Both sources of data contain key indicators around utilization of health care services and diagnoses of certain disorders, like depression, but lack the more granular race/ethnic information to understand differences across Hispanic or Latino subgroups. This work is complex, as someone's identity is related to their cultural, racial, and other social contexts that they live within. And how someone identifies may change over time as more inclusive terms or different political movements bring to light new identities.

The following questions guided this phase of CHI's research:

- Using a logistic regression approach, can a model accurately predict an individual's ethnic identity?
- Can we then apply that model to other datasets to disaggregate data across those data systems?
- What are additional research questions we can then ask once data are disaggregated within these systems?
- How can estimates of Hispanic or Latino subidentities expand opportunities to understand racial/ethnic health disparities?
- What can other data systems and states learn from this work?

In addition to the predictive modelling approach, CHI created a community outreach plan to understand the importance of data disaggregation and what conversations are needed about identity in Colorado. We spoke to key stakeholders from Hispanic or Latino communities in the state. CHI also reached out to people who are working to lower barriers to care for the Hispanic or Latino community in Colorado.

This report explains the methodological approach pursued in the second phase of this project. We have also identified lessons learned throughout the process. In addition, this report includes details from the CHAS that were integral in disaggregating data as well as the materials used for community outreach. Taken together, this report can be used as a toolkit that other systems, states, or organizations can use to pursue this research. Using this report, states can follow a framework on how to investigate the current state of data collection efforts, engage with community organizations to understand more about the specific individuals that make up their larger racial or ethnic groups, and methods to develop data collection tools to gather information on these groups.

Results have implications for overall data collection and retrospective statistical approaches by health entities. Bolstering demographic data has value for researchers and provider systems alike. The collection of social factors and other information can help connect people to additional services and programs outside of the health system that can affect overall health and well-being.

Identity in Colorado: It's Complex

Colorado has a diverse population with equally diverse needs. According to the 2021 CHAS, about 21% of residents identify as Hispanic or Latino. This represents more than 1.2 million people. Within that population, there are several cultural, ethnic, or racial identities that play a part in making up this larger ethnic monolith.

To understand this complexity, CHI expanded the ethnic identities captured by the 2021 CHAS to include specific ethnic subgroups identified by representative Hispanic- or Latino-serving community groups. CHI engaged with the following groups to understand more about groups that exist within Colorado's Hispanic or Latino community: Servicios de La Raza; Colorado Latino Leadership, Advocacy, and Research Organization; and the Latino Community Foundation of Colorado. Based on their feedback, the following identities were added to the 2021 CHAS: Caribbean, Central American, Chicano, Latinx, Mexican or Mexican American, South American, and Spanish American. An "other" option was also made available, and respondents could add their specific identity in a free response field. The ethnicity items added to the 2021 CHAS are included in Appendix A.

Ethnic identities are often developed through a complex interplay of national origin, sociopolitical context, and cultural norms. The Chicano identity, for example, emerged from a movement in California in the late 1960s that advocated for the political empowerment of Mexican Americans, through a *chicanismo* or cultural nationalism.[1] Colorado has its own historical roots in the Chicano movement due to local activism, its many agricultural regions, and Mexican or Mexican American heritage. Including Chicano as an ethnic subgroup option on the 2021 CHAS would thus give participants the opportunity to identify with a particular sociopolitical and cultural identity that is not encompassed in the broader Hispanic/Latino category.

Based on estimates from the CHAS, those who are Hispanic or Latino identify differently across subidentities (Table 1).

Table 1. Estimated Number and Percentage of Hispanic or Latino Individuals in Colorado by Subidentity, 2021*

Subidentity	Number	Percentage
Caribbean	35,636	4.1%
Central American	46,911	5.3%
Chicano	233,536	26.1%
Latinx	97,524	11.4%
Mexican/Mexican American	765,449	73.7%
Other Identity	135,522	20.4%
Spanish American	137,012	15.4%
South American	82,127	9.3%

*Respondents could choose more than one identity. Source: Colorado Health Access Survey, 2021

In Colorado, about three in four Hispanic or Latino Coloradans identify as Mexican/Mexican American, while one in four identify as Chicano. Latinx, which is a newer gender-neutral identity that typically resonates with younger generations and females, was identified by about one in 10 Hispanic or Latino individuals. These findings demonstrate that Colorado's Hispanic and Latino community represents a variety of ethnic subgroups, which can also mean unique health care needs.

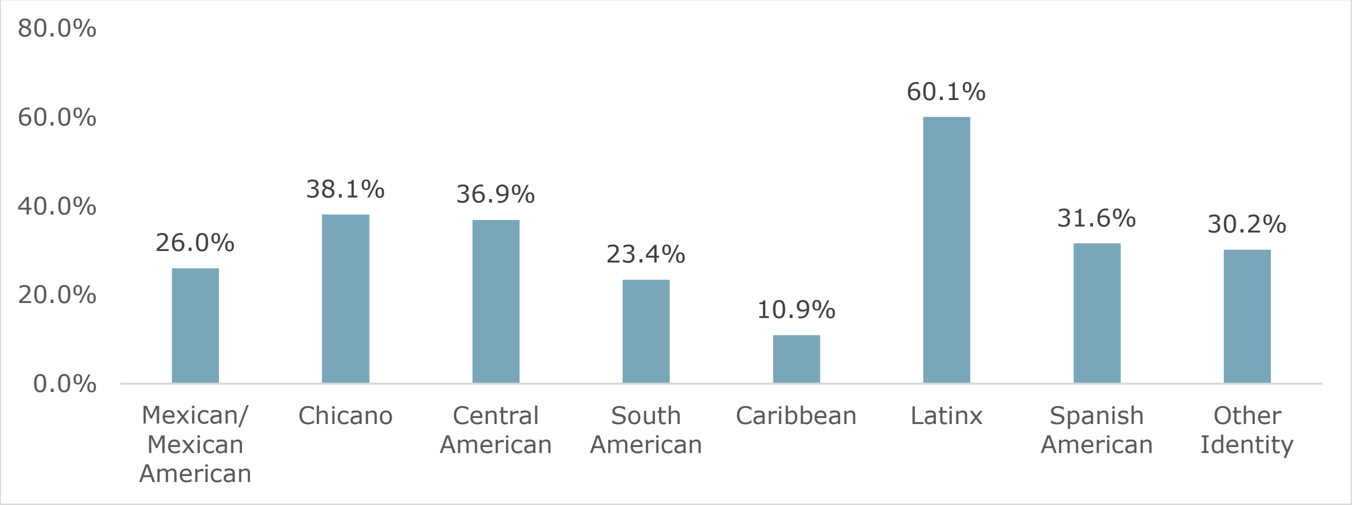
Based on CHI's conversations with community members, understanding differences within this larger Hispanic or Latino community is an important element within the health care context. Some community members said that identity and cultural aspects play a part in how people access and utilize health care services.

For example, language barriers may exist for distinct Spanish dialects, or the level of trust in the health care system may vary within subidentities, making it more difficult for members of some communities to get the care they need.

All these elements taken together create the Colorado context. An example of this context and an illustration of the importance of disaggregated data can be seen in the rates of reported poor mental health — defined in the CHAS as more than eight days when one's mental health was not good in the past month — displayed in Figure 1.

Among these subidentities, fewer Caribbean Coloradans reported poor mental health (10.9%) than Latinx Coloradans (60.1%). For comparison, the state rate of reported poor mental health in 2021 was 23.7%, while the rate for the overall Hispanic or Latino group was 25.7%.

Figure 1. Reported Poor Mental Health Rate by Hispanic or Latino Subidentities in Colorado, 2021*



*Data represents respondents ages 5 and over. Source: Colorado Health Access Survey, 2021

Such findings allow researchers to dive deeper into why differences exist across subidentities in Colorado and encourage more focused research questions to inform policy and programming. For example, what other identities — such as gender, sexuality, and race — might intersect with Hispanic or Latino subidentities to help explain these differences? Data disaggregation allows analysts and leaders to estimate key differences in their own systems. This, in turn, can create funding opportunities, expand programming, and provide services, like health navigation, to influence changes in outcomes for individuals.

While Colorado has its own ethnic context, generalizability in the methods and approach is important to the extent possible so that other states or organizations can also use data disaggregation within their own systems. With this in mind, CHI focused on characteristics like racial categories and age that are often collected in most data systems instead of other characteristics that are captured on the CHAS (specific to Colorado) and that are not usually present within other types of data systems, such as electronic health records or medical claims.

Lesson Learned

Ensure that methods are informed by specific contexts and highlight those differences so that others may try a similar approach. Geographic, cultural, social, and organizational contexts are important when shaping research questions, identifying assumptions, disseminating sharing results, and identifying how information gathered from the research can be used. However, making sure others can use similar methodological approaches is integral for testing and enhancing the developed methods.

Analysis Plan

CHI pursued a predictive modelling approach to develop regression models that could predict racial/ethnic subidentities in other data systems, and subsequently, to be used to disaggregate data. The predictive modelling approach follows four main steps: creation of the training and validation datasets, building the predictive models, evaluation of predictive model performance, and testing the models on an external data source.

CHI carried out the analysis in four steps, conducting stakeholder conversations throughout the process:

1. Data acquisition
2. Creation of training and validation datasets
3. Statistical model development
4. Model validation and results

In Step 1, CHI acquired the data necessary for the testing of the predictive models on an external data source. This step also provided the data necessary for Phase III, where CHI plans to apply the developed predictive models. In Step 2, CHI created the datasets needed for predictive modelling development and validation. In Step 3, CHI built the predictive models of interest, and, in Step 4, CHI tested the validity and predictive performance of all the predictive models developed in Step 3.

CHI identified the CHORDS and CO APCD databases as ideal datasets on which to apply the models given their comprehensiveness, their relevance to understanding how Coloradans use health care, their limited collection of race/ethnicity data currently, and the potential to gain new insights from disaggregating data. The data sources that CHI used in the analysis plan are outlined in Table 2.

Table 2. Colorado Datasets and Purpose for Inclusion in Approach

Colorado Dataset	Description	Source/Data Steward	Purpose of Inclusion
Colorado Health Access Survey (CHAS)[2]	Survey of ~10,000 Coloradans across the state	Colorado Health Institute	Collected disaggregated race/ethnicity data on the 2021 survey
Colorado All Payer Claims Database (CO APCD)[3]	Claims database from insurer groups on health care services	Center for Improving Value in Health Care	Provides data on utilization of health care services based on billing data among insured Coloradans; contains some disaggregated ethnic data
Colorado Health Observation Regional Data Service (CHORDS)[4]	Regional network of health providers along the Colorado Front Range	Regional network of health data partners	Provides data on medical services and diagnosis of health conditions from electronic health record systems

The Colorado Health Access Survey

The CHAS is Colorado’s premier source of data on health coverage, access to care, and affordability. CHI has administered the survey every other year since 2009 with the goal of providing timely information to inform policy decisions. The survey is based on a representative sample of about 10,000 randomly selected Coloradans. The first five surveys were administered solely by telephone (random digit dial), while the 2019 and 2021 surveys used an address-based sampling design in which randomly selected households receive an invitation in the mail to complete the survey online or by phone. Survey dimensions include access to care, health insurance, food insecurity, housing stability, unfair treatment in the health care system, utilization of care, behavioral health/substance use disorder, oral health, and health status. The survey is administered in English and Spanish. The CHAS has been modified numerous times to accommodate the needs and research interests of stakeholders.

The Colorado Health Observation Regional Data Service

CHORDS is a network of health systems and providers that uses electronic health record (EHR) data to identify health trends and support public health evaluation and monitoring efforts. Fourteen providers and health systems, including Kaiser Permanente, Denver Health, Children's Hospital Colorado, Clinica Family Health, STRIDE Community Health Center, and Salud Family Health Center, among others, participate as partners in the CHORDS network. The CHORDS network supports chronic disease surveillance across Colorado's counties. Currently, the CHORDS only collects aggregated data on racial or ethnic categories. These include American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, Other Race, and White.

The Colorado All-Payer Claims Database

The Center for Improving Value in Health Care (CIVHC) is a nonprofit organization that works to empower individuals, communities, and organizations through collaborative support services and health care information to advance the triple aim of better health, better care, and lower health care costs. As administrator of the CO APCD, CIVHC is steward of a comprehensive claims data set representing most insured people in Colorado and including more than 40 commercial payers, Medicaid, and Medicare. The CO APCD is a state-legislated, secure health care claims database. The complexity and scale of the database continually grows, with millions of claims submitted each month by health insurance payers representing more than 4.5 million people.

Currently, the CO APCD gathers information on aggregated racial and ethnic groups, including American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, Other Race, and White. There is some disaggregated data available that gathers information on more specific ancestral origin or ethnic identity. Based on the data extract received by CHI, these include Eastern European, Haitian, Salvadoran, Brazilian, African, African American, European, Puerto Rican, Laotian, Mexican, Mexican American, Chicano, Central American (not otherwise specified), South American (not otherwise specified), Caribbean Island, Cambodian, Cape Verdean, Vietnamese, Japanese, American, Asian Indian, Filipino, Dominican, Honduran, Columbian, Cuban, Korean, Middle Eastern, Asian, Guatemalan, Chinese, Portuguese, Russian, or Other Ethnicity. However, these data aren't universally collected, need additional quality assurance analysis, and have a lot of data missing.

Step 1: Data Acquisition

Data Overview

The first step in the methodology was to acquire information from the three data systems of interest: the CHAS, CHORDS, and CO APCD. The CHAS is a product of Colorado Health Institute, so CHI contacted the other two systems to acquire the necessary data for our approach.

IRB Approval

CHI sought Institutional Review Board (IRB) review of the research protocol before acquiring the data. CHI applied to a community IRB, Center for Research Strategies (CRS) Impact. In February 2022, CRS Impact determined the research protocol was exempt from full board review.

Applying for and Receipt of Data Extracts

Each data system had a distinct application process that CHI followed to acquire the data.

Colorado All Payer Claims Database

CHI acquired the CO APCD dataset by working with our partners at the Center for Improving Value in Health Care (CIVHC). Beginning in late October 2021, CHI engaged with CIVHC to start applying for specific elements to be included in the data extract. CHI and CIVHC determined that a de-identified extract would work best for our planned analysis and within our timeline and budget. The de-identified extract also avoided privacy issues that can arise with a more sensitive data extract.

CHI submitted the application in mid-February 2022 in preparation for the Data Release and Review Committee meeting that CHI convened with CIVHC to examine and process the data request. CHI presented the data extract submission as well as the intended use of the data to this committee. After reviewing the materials, the committee gave final approval at the beginning of March 2022.

CIVHC's data team then worked to query the data requested in the application, which took several months. CIVHC provided the data on June 29, 2022.

Colorado Health Observation Regional Data Service

CHI took a similar approach with the CHORDS dataset. CHI experts work regularly with the CHORDS network and staff, a relationship that facilitated this part of the project. With CHORDS, CHI designed an ideal list of data elements from the CHORDS database to support the project's objectives. CHI then engaged the CHORDS Research Council to refine and finalize a query of the database based on data availability, patient privacy, and governance rules. CHI began this process at the end of 2021.

CHI then met with the CHORDS Research Council for review of the project’s objectives and research questions. This required filling out a Project Intake Form through the CHORDS network to establish a use case of the CHORDS data. After review, CHI received approval for the CHORDS part of the research project in mid-March 2022.

CHI engaged with CHORDS data partners to recruit as many of them as possible to participate. In the end, CHI obtained Data Use Agreements with 10 out of the 14 data partners. This process took several months, wrapping up in June 2022.

CHI and CHORDS data programmers then designed and tested a CHORDS query on actual data. This required uncovering data or performance issues before distributing the final version of the query to all participating data partners, which were given two weeks to execute the query against their own databases. Once all queries had been returned, the University of Colorado (CU) combined the information into a single dataset. CHI provided a stipend to each data partner to acknowledge the resources applied to the project.

CU then transformed the data to ensure it was compliant with governance rules and to make it easier to use by CHI. CHI received the final data extract from CU partners on August 9, 2022.

Data Harmonization

After CHI received the CO APCD and CHORDS data, we harmonized the files to match the covariates included in the 2021 CHAS datafile. This required removing unknown or missing data and combining like-terms to create variables that match the definitions of the demographics included in the 2021 CHAS. An example of this harmonization was to create a uniform point-in-time estimate of insurance coverage. On the CHAS, insurance is created as a hierarchical variable with a single source payer. For the other two datasets, the most recent encounter or visit was used to create a single source payer to mirror the CHAS variable. To avoid any privacy or confidentiality issues, CHI stored all data files on a secured drive.

Lessons Learned

Build in sufficient time for data acquisition. Between the two datasets, data acquisition took about 10 months and over 100 hours of logged time by three CHI staff members. This was a much bigger lift than originally anticipated. Building in adequate time and resources into this part of the process is integral to sticking with proposed timelines and budget.

Build in time for additional IRB steps for data partners involved in the process. Because some data partners recently established their own IRB set up within the CHORDS governance system, they could not agree to use the community IRB that covered our project. This meant that some large health systems that are a part of the CHORDS network could not participate. Making sure that all partners are on the same page during the IRB process will build trust and help secure data from these systems.

Step 2: Creation of Training and Analysis Datasets

The predictive modelling approach starts with creating datasets to 1) develop the models (training dataset); and 2) evaluate the performance of each model (validation dataset). These datasets are typically created by randomly splitting a core dataset to avoid insertion of bias in model development.

CHI created a random 50/50 split of the 2021 CHAS dataset (50% of observations for the training dataset and 50% of observations for the validation dataset) for model development. These splits were used to provide comparable samples between the two datasets. As the sample size was small to begin with, equal sized datasets were used to make sure no bias was inserted in either sample. CHI then created separate training and validation datasets for the CHORDS and CO APCD using the CHAS. Using the CHAS allowed CHI to develop different models based on the unique characteristics of each dataset. For example, the CO APCD dataset only includes insured individuals, while CHORDS includes all individuals. The analysis approach used SAS 9.4 software.

CHI identified covariates of interest to create the predictive models and compared those to the overall 2021 CHAS dataset (see Table 3). We also analyzed descriptive statistics on the training and validation datasets to make sure that the distributions were comparable among the covariates (see Tables 4 and 5). These descriptive analyses are included in the following sections.

CHAS Analysis Samples

The 2021 CHAS datafile contained 10,405 individual records. Of those, 1,501 respondents identified as Hispanic or Latino. The CHAS dataset we used for the CHORDS analysis included all 1,501 of these respondents, while the CHAS dataset we used for the CO APCD analysis included those who were insured (1,333). CHI used these records to create the training and validation datasets that we then used to create and validate the models, outlined in the next two sections. Individuals surveyed on the CHAS could choose the Hispanic or Latino identity on either the ethnicity survey item or the race survey item. These two variables were used to create an overall Hispanic or Latino ethnicity. The survey items are available in Appendix A.

Table 3 shows a descriptive breakdown of demographic and health characteristics of the Hispanic and Latino population for each CHAS analysis compared to the total CHAS respondent sample.

Table 3. Demographic Characteristics of Analysis Samples

Demographic Characteristic	**Hispanic/Latino CHAS Analysis Sample for CHORDS** Number (Percentage)	**Hispanic/Latino CHAS Analysis Sample for CO APCD** Number (Percentage)	**Total CHAS Respondent Sample** Number (Percentage)
Total Sample	**1,501**	**1,333**	**10,405**
Subidentities[†]			
Caribbean/Central American	106 (7.1%)	86 (6.5%)	NA
Chicano	299 (19.9%)	280 (21.0%)	NA
Latinx	115 (7.7%)	102 (7.7%)	NA
Mexican/Mexican American	862 (57.4%)	744 (55.8%)	NA
South American	91 (6.1%)	81 (6.1%)	NA
Spanish American	263 (17.5%)	249 (18.7%)	NA
Missing	37 (2.5%)	36 (2.7%)	NA
Age			
Age 0 to 21	415 (27.7%)	392 (29.4%)	1,935 (18.6%)
Age 22 to 40	425 (28.3%)	352 (26.4%)	2,604 (25.0%)
Age 41 to 64	545 (36.3%)	479 (35.9%)	4,269 (41.0%)
Age 65+	101 (6.7%)	97 (7.3%)	1,491 (14.3%)
Missing	15 (1.0%)	13 (1.0%)	106 (1.0%)
Gender			
Male	700 (46.6%)	617 (46.3%)	4,992 (48.0%)
Female	791 (52.6%)	708 (53.1%)	5,326 (51.2%)
Missing	10 (0.7%)	8 (0.6%)	87 (0.8%)
Ethnicity			
Hispanic or Latino	1,501 (100%)	1,333 (100%)	1,501 (14.4%)

Not Hispanic or Latino	0 (0%)	0 (0%)	8,707 (83.7%)
Missing	0 (0%)	0 (0%)	197 (1.9%)
Race*			
American Indian or Alaska Native	104 (6.9%)	95 (7.1%)	270 (2.6%)
Asian	16 (1.1%)	15 (1.1%)	292 (2.8%)
Black or African American	44 (2.9%)	42 (3.2%)	506 (4.9%)
Some Other Race***	57 (3.8%)	53 (4.0%)	307 (3.0%)
White	534 (35.6%)	493 (37.0%)	8,387 (80.6%)
Missing (Any Race Data)	21 (1.4%)	19 (1.4%)	289 (2.8%)
Preferred Language Spoken at Home			
Speaks a Language Other Than English at Home	683 (45.5%)	567 (42.5%)	1,179 (11.3%)
English Spoken at Home	814 (54.2%)	763 (57.2%)	9,148 (87.9%)
Missing	4 (0.3%)	3 (0.2%)	78 (0.8%)
Insurance Coverage Type**			
Private Insurance	806 (53.7%)	806 (60.5%)	6,676 (64.2%)
Medicaid/Child Health Plan *Plus* (CHP+)	434 (28.9%)	434 (32.6%)	1,703 (16.4%)
Medicare	83 (5.5%)	83 (6.2%)	1,377 (13.2%)
Other Insurance	10 (0.7%)	10 (0.8%)	69 (0.7%)
Uninsured	168 (11.2%)	NA	580 (5.6%)
Missing	0 (0%)	0 (0%)	0 (0%)
Rurality			
Lives in a Rural County	518 (34.5%)	467 (35.0%)	4,069 (39.1%)

Lives in Urban County	983 (65.5%)	866 (65.0%)	6,336 (60.9%)
Missing	0 (0%)	0 (0%)	0 (0%)

*Race categories were non-mutually exclusive. This percentage represents the additional race groups that the Hispanic or Latino respondent also denoted as an identity for the two analysis samples. **Insurance coverage types were mutually exclusive. ***Some Other Race includes Native Hawaiian or Other Pacific Islander, Middle Eastern or North African, or some other race reported. † Subidentities were not included for the total sample, since it was only collected for those who identified as Hispanic or Latino. Source: 2021 Colorado Health Access Survey.

The Hispanic or Latino population in CHAS has distinct differences compared with the overall CHAS data sample. The sample was more likely to be under age 19, also identify as American Indian or Alaska Native, be uninsured, and speak a language other than English at home. Because of small sample sizes for Caribbean and Central American subidentities, these two identities were aggregated into a new model for inclusion in the analysis.

Using the CHAS for CHORDS Model Development

Of the 1,501 respondents included in the CHAS datafile used for the CHORDS model development, 751 were randomly assigned to the training dataset, while 750 were assigned to the validation dataset. CHI analyzed the descriptive statistics of both samples to understand the comparability of the training and validation datasets before model development. Those results are provided in Table 4.

Table 4. Demographic Characteristics of Training and Validation Datasets Used in the Model Development for the CHORDS

Demographic Characteristic	**Training Dataset** Number (Percentage) *(Confidence Interval)*	**Validation Dataset** Number (Percentage) *(Confidence Interval)*
Total Sample	**751**	**750**
Subidentities		
Caribbean/Central American	52 (6.9%) *(5.1%, 8.7%)*	54 (7.2%) *(5.3%, 9.1%)*
Chicano	151 (20.1%) *(17.2%, 23.0%)*	148 (19.7%) *(16.9%, 22.6%)*
Latinx	66 (8.8%) *(6.8%, 10.8%)*	49 (6.5%) *(4.8%, 8.3%)*
Mexican/Mexican American	437 (58.2%)	425 (56.7%)

	(54.7%, 61.7%)	*(53.1%, 60.2%)*
South American	46 (6.1%) *(4.4%, 7.8%)*	45 (6.0%) *(4.3%, 7.7%)*
Spanish American	129 (17.2%) *(14.5%, 19.9%)*	134 (17.9%) *(15.1%, 20.6%)*
Missing (All Subidentity Data)	24 (3.2%) *(1.9%, 4.5%)*	13 (1.7%) *(0.8%, 2.7%)*
Age		
Age 0 to 21	191 (25.4%) *(22.3%, 28.6%)*	224 (29.9%) *(26.6%, 33.1%)*
Age 22 to 40	213 (28.4%) *(25.1%, 31.6%)*	212 (28.3%) *(25.0%, 31.5%)*
Age 40 to 64	280 (37.3%) *(33.8%, 40.7%)*	265 (35.3%) *(31.9%, 38.8%)*
Age 65+	57 (7.6%) *(5.7%, 9.5%)*	44 (5.9%) *(4.2%, 7.6%)*
Missing	10 (1.3%) (0.5%, 2.2%)	5 (0.7%) (0.1%, 1.3%)
Gender		
Male	334 (44.5%) *(40.9%, 48.0%)*	366 (48.8%) *(45.2%, 52.4%)*
Female	413 (55.0%) *(51.4%, 58.6%)*	378 (50.4%) *(46.8%, 54.0%)*
Missing	4 (0.5%) *(0.0%, 1.1%)*	6 (0.8%) *(0.2%, 1.4%)*
Race*		
American Indian or Alaska Native	49 (6.5%) *(4.8%, 8.3%)*	55 (7.3%) *(5.5%, 9.2%)*
Asian	8 (1.1%) *(0.3%, 1.8%)*	8 (1.1%) *(0.3%, 1.8%)*
Black or African American	19 (2.5%)	25 (3.3%)

	(1.4%, 3.7%)	*(2.0%, 4.6%)*
Some Other Race***	32 (4.3%) *(2.8%, 5.7%)*	25 (3.3%) *(2.0%, 4.6%)*
White	269 (35.8%) *(32.4%, 39.3%)*	265 (35.3%) *(31.9%, 38.8%)*
Missing	11 (1.5%) *(0.6%, 2.3%)*	10 (1.3%) *(0.5%, 2.2%)*
Preferred Language Spoken at Home		
Speaks a Language Other Than English at Home	348 (46.3%) *(42.8%, 49.9%)*	335 (44.7%) *(41.1%, 48.2%)*
English Spoken at Home	402 (53.5%) *(50.0%, 57.1%)*	412 (54.9%) *(51.4%, 58.5%)*
Missing	1 (0.1%) *(0.0%, 0.4%)*	3 (0.4%) *(0.0%, 0.9%)*
Insurance Coverage Type**		
Private/Other Insurance	399 (53.1%) *(49.6%, 56.7%)*	417 (55.6%) *(52.0%, 59.2%)*
Public Insurance (Medicare, Medicaid, Child Health Plan *Plus* (CHP+))	270 (36.0%) *(32.5%, 39.4%)*	247 (32.9%) *(29.6%, 36.3%)*
Uninsured	82 (10.9%) *(8.7%, 13.2%)*	86 (11.5%) *(9.2%, 13.8%)*
Missing	0 (0%)	0 (0%)
Rurality		
Lives in a Rural County	252 (33.6%) *(30.2%, 36.9%)*	266 (35.5%) *(32.0%, 38.9%)*
Lives in an Urban County	499 (66.4%) *(63.1%, 69.8%)*	484 (64.5%) *(61.1%, 68.0%)*
Missing	0 (0%)	0 (0%)

*Race categories were non-mutually exclusive. This percentage represents the additional race groups that the Hispanic or Latino respondent also denoted as an identity. **Insurance coverage types were mutually exclusive. ***Some Other Race includes Native Hawaiian or Other Pacific Islander, Middle Eastern or North African, or some other race reported.

Based on the confidence intervals of the estimates, there are no statistically significant differences in the estimates of the covariates between the training and validation datasets. CHI pursued model development steps for the CHAS dataset created for the CHORDS models using these datasets. Because of small estimates, CHI removed the covariate of Asian race from the model development analysis.

Using the CHAS for CO APCD Model Development

Of the 1,333 respondents included in the CHAS datafile used for the CO APCD model development, 667 were randomly assigned to the training dataset, while 666 were assigned to the validation dataset. CHI analyzed the descriptive statistics of both samples to understand the comparability of the training and validation datasets before model development. Those results are provided in Table 5.

Based on the confidence intervals of the estimates, there are no statistically significant differences in the estimates of the covariates between the training and validation datasets. CHI pursued model development steps for the CHAS dataset created for the CO APCD models using these datasets. Because of small estimates, CHI removed the covariate of Asian race from the model development analysis.

Table 5. Demographic Characteristics of Training and Validation Datasets Used in the Model Development for the CO APCD

Demographic Characteristic	**Training Dataset** Number (Percentage) *(Confidence Interval)*	**Validation Dataset** Number (Percentage) *(Confidence Interval)*
Total Sample	**667**	**666**
Subidentities		
Caribbean/Central American	45 (6.7%) *(4.8%, 8.7%)*	41 (6.2%) *(4.3%, 8.0%)*
Chicano	134 (20.1%) *(17.0%, 23.1%)*	146 (21.9%) *(18.8%, 25.1%)*
Latinx	54 (8.1%) *(6.0%, 10.2%)*	48 (7.2%) *(5.2%, 9.2%)*
Mexican/Mexican American	370 (55.5%) *(51.7%, 59.3%)*	374 (56.2%) *(52.4%, 59.9%)*
South American	37 (5.5%) *(3.8%, 7.3%)*	44 (6.6%) *(4.7%, 8.5%)*
Spanish American	111 (16.6%)	138 (20.7%)

	(13.8%, 19.5%)	*(17.6%, 23.8%)*
Missing (All Subidentity Data)	17 (2.5%) *(1.3%, 3.7%)*	19 (2.9%) *(1.6%, 4.1%)*
Age		
Age 0 to 21	193 (28.9%) *(25.5%, 32.4%)*	199 (29.9%) *(26.4%, 33.4%)*
Age 22 to 40	173 (25.9%) *(22.6%, 29.3%)*	179 (26.9%) *(23.5%, 30.3%)*
Age 40 to 64	240 (36.0%) *(32.3%, 39.6%)*	239 (35.9%) *(32.2%, 39.5%)*
Age 65+	52 (7.8%) *(5.8%, 9.8%)*	45 (6.8%) *(4.8%, 8.7%)*
Missing	9 (1.3%) *(0.5%, 2.2%)*	4 (0.6%) *(0.0%, 1.2%)*
Gender		
Male	307 (46.0%) *(42.2%, 49.8%)*	310 (46.5%) *(42.7%, 50.3%)*
Female	356 (53.4%) *(49.6%, 57.2%)*	352 (52.9%) *(49.1%, 56.7%)*
Missing	4 (0.6%) *(0.0%, 1.2%)*	4 (0.6%) *(0.0%, 1.2%)*
Race*		
American Indian or Alaska Native	49 (7.3%) *(5.4%, 9.3%)*	46 (6.9%) *(5.0%, 8.8%)*
Asian	4 (0.6%) *(0.0%, 1.2%)*	11 (1.7%) *(0.7%, 2.6%)*
Black or African American	23 (3.4%) *(2.1%, 4.8%)*	19 (2.9%) *(1.6%, 4.1%)*
Some Other Race***	22 (3.3%) *(1.9%, 4.7%)*	31 (4.7%) *(3.1%, 6.3%)*
White	241 (36.1%) *(32.5%, 39.8%)*	252 (37.8%) *(34.1%, 41.5%)*

Missing (All Race Data)	14 (2.1%) *(1.0%, 3.2%)*	5 (0.8%) *(0.1%, 1.4%)*
Preferred Language Spoken at Home		
Speaks a Language Other Than English at Home	281 (42.1%) *(38.4%, 45.9%)*	286 (42.9%) *(39.2%, 46.7%)*
English Spoken at Home	385 (57.8%) *(54.0%, 61.5%)*	378 (56.8%) *(53.0%, 60.5%)*
Missing	1 (0.1%) *(0.0%, 0.4%)*	2 (0.2%) *(0.0%, 0.7%)*
Insurance Coverage Type**		
Private/Other Insurance	403 (60.4%) *(56.7%, 64.1%)*	413 (62.0%) *(58.3%, 65.7%)*
Public Insurance (Medicare, Medicaid, Child Health Plan *Plus* (CHP+))	264 (39.6%) *(35.9%, 43.3%)*	253 (38.0%) *(34.3%, 41.7%)*
Uninsured	NA	NA
Missing	0 (0%)	0 (0%)
Rurality		
Lives in a Rural County	226 (33.9%) *(30.3%, 37.5%)*	241 (36.2%) *(32.5%, 39.8%)*
Lives in an Urban County	441 (66.1%) *(62.5%, 69.7%)*	425 (63.8%) *(60.2%, 67.5%)*
Missing	0 (0%)	0 (0%)

*Race categories were non-mutually exclusive. This percentage represents the additional race groups that the Hispanic or Latino respondent also denoted as an identity. **Insurance coverage types were mutually exclusive. ***Some Other Race includes Native Hawaiian or Other Pacific Islander, Middle Eastern or North African, or some other race reported.

Lessons Learned

Verify data storage capacity before requesting datasets. Data acquired from CIVHC for the CO APCD was over 43 gigabytes. The time it takes to download these files must be considered when acquiring these data, as it could take several workdays just to download them onto servers. CHI also has a storage capacity of 1 terabyte on our secure data servers, which provided enough storage capacity for the CHORDS and CO APCD datasets. These two issues can greatly impact an organization's ability to download and store massive datafiles. This will also impact computational processing time. The larger the files, the longer it will take for the available technology being used for analysis to process the data.

Capitalize on existing data disaggregation efforts. A part of the community and data steward engagement should include a scan of current state of data initiatives. Institutions and organizations understand the importance of specific data and some partners might already be gathering granular data. Even when disaggregated data are not available on all records, the limited disaggregated information can be used in the validation of models.

Step 3. Statistical Model Development

Testing of Covariates

After creation and analysis of the training and validation datasets, CHI then developed the predictive models. CHI modeled each Hispanic or Latino subidentity as a dichotomous outcome in its own predictive model. Those who identified as Caribbean or Central American were aggregated due to small sample sizes. We designed predictive models for the following subidentities as the outcome of interest:

- Caribbean/Central American
- Chicano
- Latinx
- Mexican/Mexican American
- South American
- Spanish American

CHI investigated comparable demographic factors across the three datasets. Table 6 outlines those that are included across the three data systems and could be used as covariates within the logistic models.

Table 6. Variables Investigated as Covariates in Predictive Modelling Analysis

Common Variables Across All Systems	Additional Common Variables Between CHAS and CO APCD	Additional Common Variables Between CHAS and CHORDS
Race, Age, Gender, Rurality, Insurance Coverage Type	Disaggregated country of origin/ethnicity	Language spoken at home

Coding of the Covariates

CHI categorized the covariates for the logistic regression modelling approach. Those codes are provided in Table 7. Variables were kept consistent across all models investigated. The one exception the age covariate. Sample size became an issue when comparing younger age groups to the age 65 and older population. In those instances, the age covariate was kept continuous.

The covariate variable names are used throughout the rest of the report for simplicity. CHI chose reference groups based on the "No" or "0" for dichotomous outcomes where applicable for simplicity in coding. For insurance coverage, private/other insurance was chosen as the reference category, while the age group 41 to 64 was chosen for the age categorization variable. Use of the CLASS option in PROC LOGISTIC creates dummy codes for all variables. The reference group noted becomes "0" during this operation.

Table 7. Codes for Modelling Approach and Reference Group

Covariate Name	Codes	Reference Group
Age*	**1** = Age 0 to 21, **2** = Age 22 to 40, **3** = Age 41 to 64, **4** = Age 65+	**3**
American Indian or Alaska Native	**1** = Yes, **0** = No	**0**
Black or African American	**1** = Yes, **0** = No	**0**
Gender	**1** = Male, **0** = Female	**0**
Insurance Coverage Type	**1** = Private/Other Insurance, **2** = Public Insurance (Medicare, Medicaid, CHP+), **3** = Uninsured	**1**
Other Language Spoken at Home	**1** = Yes, Speaks Another Language Than English at Home, **0** = No, Speaks English at Home	**0**
Some Other Race	**1** = Yes, **0** = No	**0**
Rurality	**1** = Urban County, **0** = Rural County	**0**
White	**1** = Yes, **0** = No	**0**

*Exception was used for the Latinx model (both data sources) and the Caribbean/Central American model developed for the CHORDS. Those who identified as Latinx were much more likely to be younger. The Caribbean/Central American showed issues for sample size across the different age groups. These models used a continuous age variable.

Bivariate Analysis and Stepwise Logistic Regression

CHI completed a bivariate analysis to test for a statistically significant relationship between the covariate and the subidentity. We established a threshold of 0.2 as a significance level to keep the variable in for stepwise model fitness testing. Modelling techniques use a threshold between 0.15 to 0.25, so CHI decided on a 0.2 threshold within this range.[5]

Table 8 displays the results of the bivariate analyses. We developed two sets of models based on the data system. The CHORDS data, because of the presence of a language variable, had an expanded model that included language as a covariate.

We used a stepwise selection modelling approach to understand how well each additional covariate increased the fitness of the regression model. This allowed us to evaluate model fitness with each added covariate. Increased model fitness refers to how well the observed data correspond to the fitted model developed. CHI applied the -log likelihood value to understand if each added covariate increased the model fitness, using a significance threshold of 0.05 to measure if the added covariate was kept in the model based on the -log likelihood p value.

Understanding the effectiveness of the model is described by the model's ability to discriminate and how well it is calibrated. We used the concordance (or c) statistic to understand if each additional metric contributed to overall model fitness. The c statistic is used to assess a logistic model's discrimination ability, which refers to the model's ability to distinguish between those with and those without the outcome of interest. The predictive power of the model increases as this statistic increases.[6]

To understand the calibration of the logistic model, CHI ran a Hosmer-Lemeshow (H/L) test on each subidentity model. This test measures the comparison between the observed and the expected outcome of each subidentity. We chose this test because it is more suitable for models with lower sample sizes.[7,8] A threshold of 0.05 was used as a threshold for the H/L test. Smaller p values for this test show that the model's fitness should be reevaluated.

If these fitness statistics were favorable with addition of a new covariate, it was kept in the model.

Table 8. Univariate Analysis of Covariates with Subidentity as Outcome of Interest for CHORDS and CO APCD Datasets

Subidentity Model	**CHORDS Training Dataset** Covariate (P Value)	**CO APCD Training Dataset*** Covariate (P Value)
Caribbean/ Central American	Age (0.25) Gender (0.98) Insurance Coverage Type (0.03) Other Language Spoken at Home (<0.01) American Indian or Alaska Native (0.23) Black or African American (0.72)	Age (0.04) Gender (0.19) Insurance Coverage Type (0.16) American Indian or Alaska Native (0.24) Black or African American (0.07) Other Race (0.73)

	Other Race (0.37) White (0.03) Rurality (0.98)	White (0.88) Rurality (0.37)
Chicano	Age (0.31) Gender (<0.01) Insurance Coverage Type (<0.01) Other Language Spoken at Home (0.01) American Indian or Alaska Native (<0.01) Black or African American (0.58) Other Race (0.76) White (0.05) Rurality (0.66)	Age (0.71) Gender (0.03) Insurance Coverage Type (<0.01) American Indian or Alaska Native (0.10) Black or African American (0.75) Other Race (0.07) White (0.15) Rurality (0.40)
Latinx	Age (<0.01) Gender (0.66) Insurance Coverage Type (0.04) Other Language Spoken at Home (0.12) American Indian or Alaska Native (0.31) Black or African American (0.92) Other Race (0.47) White (0.06) Rurality (0.50)	Age (<0.01) Gender (0.62) Insurance Coverage Type (0.12) American Indian or Alaska Native (0.44) Black or African American (0.51) Other Race (0.49) White (0.89) Rurality (0.91)
Mexican/ Mexican American	Age (0.10) Gender (0.40) Insurance Coverage Type (0.41) Other Language Spoken at Home (0.05) American Indian or Alaska Native (0.15) Black or African American (0.04) Other Race (0.27) White (0.11) Rurality (0.11)	Age (0.11) Gender (0.40) Insurance Coverage Type (0.16) American Indian or Alaska Native (0.07) Black or African American (<0.01) Other Race (0.66) White (0.06) Rurality (0.12)
South American	Age (0.43) Gender (0.64) Insurance Coverage Type (<0.01) Other Language Spoken at Home (<0.01) American Indian or Alaska Native (0.62) Black or African American (0.43) Other Race (0.46) White (0.14) Rurality (0.05)	Age (0.99) Gender (0.96) Insurance Coverage Type (<0.01) American Indian or Alaska Native (0.69) Black or African American (0.03) Other Race (0.98) White (0.03) Rurality (0.13)

Spanish American	Age (0.02) Gender (0.35) Insurance Coverage Type (0.02) Other Language Spoken at Home (<0.01) American Indian or Alaska Native (<0.01) Black or African American (0.90) Other Race (0.45) White (<0.01) Rurality (<0.01)	Age (0.01) Gender (0.78) Insurance Coverage Type (<0.01) American Indian or Alaska Native (<0.01) Black or African American (0.29) Other Race (0.92) White (<0.01) Rurality (<0.01)

*Language was not available as a covariate within the CO APCD dataset and was excluded from model analysis for the CO APCD model development. Source: 2021 Colorado Health Access Survey

We found a variety of covariates to be the best predictors for each model. These also created the best fitness scores. Table 9 displays these covariates for the CHORDS models and Table 10 displays them for the CO APCD models.

Table 9. Covariates Included in Each Subidentity Logistic Regression Model and Model Fitness Statistics Developed for the CHORDS

Subidentity Model	**Covariates In Model**	**Model Fitness Statistics** (Training)*	**Model Fitness Statistics** (Validation)*
Caribbean/Central American	American Indian or Alaska Native, Black or African American, White, Other Language Spoken at Home, Age (Continuous), Insurance Coverage, Gender	**C:** 0.70 **H/L:** 11.7 (0.16)	**C:** 0.69 **H/L:** 7.9 (0.45)
Chicano	American Indian or Alaska Native, White, Age, Other Language Spoken at Home, Gender, Rurality, Insurance Coverage Type	**C:** 0.69 **H/L:** 6.3 (0.61)	**C:** 0.66 **H/L:** 2.6 (0.96)
Latinx	American Indian or Alaska Native, White, Age (Continuous), Gender, Other Language Spoken at Home, Insurance Coverage, Rurality	**C:** 0.69 **H/L:** 4.7 (0.79)	**C:** 0.61 **H/L:** 3.6 (0.89)
Mexican/Mexican American	American Indian or Alaska Native, Black or African American, White, Age, Rurality, Other Language Spoken at Home, Insurance Coverage Type	**C:** 0.61 **H/L:** 6.6 (0.58)	**C:** 0.66 **H/L:** 3.7 (0.89)

South American	White, Other Language Spoken at Home, Rurality, Insurance Coverage Type, Age, Rurality	**C:** 0.77 **H/L:** 2.7 (0.95)	**C:** 0.70 **H/L:** 2.6 (0.95)
Spanish American	American Indian or Alaska Native, White, Other Language Spoken at Home, Age, Insurance Coverage Type, Rurality	**C:** 0.74 **H/L:** 10.7 (0.22)	**C:** 0.74 **H/L:** 10.9 (0.21)

*H/L: Hosmer-Lemeshow Goodness-of-Fit Test Statistics, Chi-square value (P-value); C: C Statistic

Table 10. Covariates Included in Each Subidentity Logistic Regression Model and Model Fitness Statistics Developed for the CO APCD

Subidentity Model	**Covariates In Model**	**Model Fitness Statistics** (Training)*	**Model Fitness Statistics** (Validation)*
Caribbean/Central American	American Indian or Alaska Native, Black or African American, Age, Insurance Coverage Type, Gender, Rurality	**C:** 0.68 **H/L:** 5.2 (0.63)	**C:** 0.60 **H/L:** 3.8 (0.87)
Chicano	American Indian or Alaska Native, White, Some Other Race, Age, Gender, Rurality, Insurance Coverage Type	**C:** 0.65 **H/L:** 8.0 (0.44)	**C:** 0.64 **H/L:** 3.3 (0.91)
Latinx	American Indian or Alaska Native, Age (Continuous), Gender, Insurance Coverage Type	**C:** 0.66 **H/L:** 5.9 (0.66)	**C:** 0.58 **H/L:** 8.2 (0.42)
Mexican/Mexican American	American Indian or Alaska Native, Black or African American, White, Age, Rurality, Insurance Coverage Type	**C:** 0.63 **H/L:** 8.3 (0.41)	**C:** 0.62 **H/L:** 6.1 (0.63)
South American	American Indian or Alaska Native, White, Rurality, Age, Insurance Coverage Type, Age	**C:** 0.71 **H/L:** 8.8 (0.36)	**C:** 0.68 **H/L:** 8.8 (0.45)
Spanish American	American Indian Alaska Native, White, Age, Insurance Coverage Type, Rurality	**C:** 0.71 **H/L:** 6.1 (0.64)	**C:** 0.67 **H/L:** 9.6 (0.22)

*H/L: Hosmer-Lemeshow Goodness-of-Fit Test Statistics, Chi-square value (P-value); C: C Statistic

The optimum c statistic is 1. A c statistic value of 1 indicates that the model is correctly identifying all those who were the outcome of interest (identifying those who were Spanish American, for example) and correctly identifying all of those who were not the outcome of interest (identifying those who were not Spanish American, for example). This is describing both the sensitivity and specificity of the model output, which will be discussed more in depth in the next section.

Step 4: Model Validation and Results

Validation of the Logistic Regression Models from the CHAS

After CHI developed the subidentity models and found optimized model fitness statistics, we analyzed the sensitivity and specificity on each logistic regression model for the CHAS models built for both the CHORDS and CO APCD.

To do this, CHI applied the models developed on the training datasets to the validation datasets. We created a receiver operating curve (ROC) for each model to understand how well they discriminated between each record to predict the subidentity of interest. To understand performance, the ROC analysis identifies how well the model detects that someone is the subidentity of interest and how accurately it rules out someone who is not that subidentity. The area under the curve (AUC) statistic is used to understand the discriminating power of the model. In logistic regression, the AUC statistic and c statistic equal the same value.

The further away the AUC value is from 0.5, which represents random chance of the outcome occurring, the better the model is at predicting the outcome of interest. In other words, the closer the AUC value gets to 1, the better the predictive performance of the model.[9,10] The ROC analysis also determines the predictive probabilities of each model and can identify the optimum threshold where the model is performing at its peak. The predicted probabilities are calculated using the coefficients in each model for the subidentities by transforming the logistic regression function, described by the following formula:

$$P\ (t) = \frac{1}{1+ e^{-t}}$$

The optimum predictive threshold also helps identify the optimum sensitivity and specificity of the model. A sensitivity analysis, also called the true positive rate, measures the proportion of actual positives (someone who is the subidentity) that were correctly identified. The specificity analysis measures the proportion of actual negatives (someone who is not the subidentity) that were correctly identified. The method for identifying the optimum predictive threshold was derived from the calculation of the distance to the optimum point on the ROC analysis (0,1), where 1-specificity = 0 and sensitivity = 1. That value is found from taking the square root of (1-sensitivity)^2 + (1-specificity)^2.[11] Table 11 displays the result of the ROC analysis. CHI generated ROC curves for each model for the corresponding data source of interest. Those curves are available in Appendix B.

Table 11. Results from ROC Curve Analysis: Sensitivity and Specificity Results, Validation Datasets

Subidentity Model	**Validation Statistics** (CHORDS)	**Validation Statistics** (CO APCD)
Caribbean/Central American	**AUC:** *0.69* **Optimum Threshold:** 0.09 Sensitivity: 68.5% Specificity: 59.4%	**AUC:** *0.60* **Optimum Threshold:** 0.09 Sensitivity: 53.7% Specificity: 64.7%
Chicano	**AUC:** *0.66* **Optimum Threshold:** 0.26 Sensitivity: 59.4% Specificity: 63.3%	**AUC:** *0.64* **Optimum Threshold:** 0.27 Sensitivity: 62.9% Specificity: 57.8%
Latinx	**AUC:** *0.61* **Optimum Threshold:** 0.09 Sensitivity: 71.1% Specificity: 49.4%	**AUC:** *0.58* **Optimum Threshold:** 0.11 Sensitivity: 51.1% Specificity: 70.6%
Mexican/Mexican American	**AUC:** *0.66* **Optimum Threshold:** 0.65 Sensitivity: 62.3% Specificity: 62.9%	**AUC:** *0.62* **Optimum Threshold:** 0.68 Sensitivity: 59.0% Specificity: 59.9%
South American	**AUC:** *0.70* **Optimum Threshold:** 0.10 Sensitivity: 62.2% Specificity: 67.7%	**AUC:** *0.68* **Optimum Threshold:** 0.10 Sensitivity: 77.3% Specificity: 58.0%
Spanish American	**AUC:** *0.74* **Optimum Threshold:** 0.25 Sensitivity: 70.0% Specificity: 66.7%	**AUC:** *0.67* **Optimum Threshold:** 0.26 Sensitivity: 64.2% Specificity: 62.7%

Source: 2021 Colorado Health Access Survey

The farther away the AUC value is from 0.5 (which denotes no discrimination is occurring), the better the model is performing. This is because a value of 0.5 suggests that the model is no better at predicting the outcome than it would perform by just random chance. The rule of thumb is that an AUC value of 0.5-0.7 represents poor discrimination, 0.7-0.8 is acceptable discrimination, 0.8-0.9 is excellent discrimination, and over 0.9 is outstanding discrimination of the model.[12]

Based on this threshold rule, the CHAS models built for the CHORDS dataset with better performance are Caribbean/Central American, Spanish American, and South American. The best performing CHAS models built for the CO APCD dataset were South American and Spanish American subidentities. These two models, though, had poor discrimination.

Testing of the CHAS Logistic Regression Models Using the CO APCD

CHI performed a test on an external data source to understand the performance of the models in a real-world application of the models. In the CO APCD dataset, some disaggregated identities are captured on a subset of the data. These data are not uniformly gathered across payers, as they are not mandatory reporting fields, and the data are missing on about 8% of individual records. In addition, over 60% of the sample had a disaggregated subidentity labeled as “other,” which provides no specifics or additional information about the individual.

CHI used these disaggregated data to understand and test the model predictability for the few identities that had available data. These included Caribbean/Central American identities and South American identities.

To create the dataset for testing, CHI removed all observations where respondents did not identify as Hispanic or Latino. This limited the CO APCD record count to 686,671. We also excluded records that were missing disaggregated ethnicity data, which brought the final observation count to 531,957.

CHI then categorized those with data that had a reported code associated with an identity or country of origin from Central America or the Caribbean. The same approach was used for those with data associated with South America. Only 36 individuals (0.007%) of the total sample were categorized as Central American or Caribbean. This was similar for those categorized as South American — only 95 individuals (0.018%) fell into this category. The sample may not be representative of the Colorado population who identify as these ethnicities.

Based on the CHAS sample built for the CO APCD, about 7% of the Hispanic/Latino population identify as either Central American or Caribbean, while 6% identify as South American. In addition, sample size has an impact on predictive modelling approaches and the ability to discern outcomes. The small counts available in the disaggregated data fields could reduce the predictive power of the models and insert bias in the maximum likelihood estimation in logistic regression.[13]

CHI applied the models to the datasets, and then performed sensitivity and specificity analyses to understand the two models’ abilities to discern the true outcomes. Those results are presented in Table 12.

Based on the models, 77,253 individuals (14.5%) were identified as Caribbean/Central American, while 109,961 individuals (20.7%) were identified as South American, using the optimum thresholds calculated. The South American model had a higher specificity (79.3%), meaning there was fewer false positives categorized as this identity when the individual was not South American. However, the sensitivity was low (17.9%), meaning there were many false negatives. The Caribbean/Central American model provided a sensitivity estimate of 41.7% and a specificity of 85.5%.

In this scenario, because of the small sample size of each identity within the CO APCD sample, the high rate of specificity calculated is very important. Minimizing the number of positives shows that the model can discern between those who are not the outcome of interest.

Table 12. Specificity and Sensitivity Analysis of the CO APCD Dataset with Disaggregated Race Data

Subidentity Model	**Number of Observations Flagged by Model** (Percentage of Total Sample)	**Sensitivity**	**Specificity**
Caribbean/Central American	77,253 (14.5%)	41.7%	85.5%
South American	109,961 (20.7%)	17.9%	79.3%

Source: Colorado All Payer Claims Database, 2016-2022

There may also be a limitation of the data source, as the disaggregated ethnicity data is not uniformly collected and has not been validated as much as the aggregated race groups. Additionally, because the representation of these ethnicities was substantially smaller in the disaggregated CO APCD data compared to the CHAS, the model could be identifying individuals who might have otherwise identified as South American or Central American/Caribbean.

Also, sample size is always an issue when using logistic regression. Because of the small number of observations that had disaggregated ethnicity data available, and the presence of less than 0.01% of each identity present within that sample, this inserts issues in the predictive model performing as well.

Although there were limitations to using the disaggregated data available on the CO APCD, the validation process was still valuable in understanding more about the developed models. Because of issues with sample size and representativeness of the data available, more validation testing from outside sources would be beneficial.

Results of the Model Validation Process

Across the six models, South American, Spanish American, and Caribbean/Central American had the best performing models developed to be used on the CHORDS. The South American and Spanish American models developed to be used on the CO APCD had the best performance. Upon applying the Caribbean/Central American model to the available disaggregated data in the CO APCD, we found that this model performed well to identity those who were not the Caribbean or Central American identity and identified over 40% of those who did identify as Caribbean or Central American within the dataset.

Model performance differed between the models developed for the CHORDS dataset and the CO ACPD dataset. The key difference was the inclusion of language as a covariate; CHORDS included language and the CO APCD did not. This difference highlights the importance of understanding whether someone's preferred language is something other than English, which increases the model's ability to discriminate the outcome of interest. The importance of including language in a model also makes sense from a sociological perspective, as someone's language has a profound impact on a person's cultural identity as well as how well the person can understand and traverse the health care system.

It also makes sense in the Colorado context: Nearly 1 million people speak a language other than English at home, which relates to someone's ethnic heritage.[14,15] The sociocultural importance of language is also demonstrated statistically in the predictive models.

Based on the CHAS, about three in four individuals in Colorado identified as Mexican or Mexican American in 2021. This aligns with what CHI also concluded from conversations with community members. The fact that the Mexican or Mexican American subidentity model did not perform well may point to how there isn't enough information available to statistically differentiate this group within the dataset. As a large proportion of individuals identified as Mexican or Mexican American within the overall dataset used for model development, additional covariates may be required to differentiate between someone who does and does not identify as Mexican or Mexican American.

Table 13 and Table 14 display the coefficients and logistic regression equations. The in-depth predictive modelling output from the PROC LOGISTIC procedures for each model are provided in Appendix C.

Table 13. Logistic Regression Equations for CHAS Models Developed for the CHORDS

Subidentity Model	Coefficients and Logistic Regression Equations	Calculated Threshold
Caribbean/Central American*	**Logit(P) =** -2.3724 – 0.4269 (American Indian or Alaska Native) + 0.0207 (White) + 1.5247 (Black or African American) + 1.26 (Other Language Spoken at Home) – 0.0141 (Age Continuous) – 0.4137 (Public Insurance) – 0.3664 (Uninsured)	0.08733029 85808655
Chicano	**Logit(P) =** -0.571 + 0.0487 (American Indian or Alaska Native) – 0.6174 (White) – 0.2303 (Age 0 to 21) + 0.0884 (Age 22 to 40) + 0.2964 (Age 65+) – 0.5499 (Other Language Spoken at Home) + 0.013 (Male) – 0.446 (Urban) + 0.6793 (Public Insurance) – 0.3197 (Uninsured)	0.25998977 3078588
Latinx*	**Logit(P) =** -2.9072 + 0.6688 (American Indian or Alaska Native) + 0.066 (White) – 0.0031 (Age Continuous) – 0.2521 (Other Language Spoken at Home) + 0.1141 (Public Insurance) + 0.079 (Uninsured) + 0.768 (Urban)	0.08867773 14132146
Mexican/Mexican American	**Logit(P) =** -0.2228 – 0.2429 (American Indian or Alaska Native) – 0.604 (Black or African American) – 0.0784 (White) + 0.7151 (Age 0 to 21) + 0.4056 (Age 22 to 40) + 0.5501 (Age 65+) + 0.641 (Other Language Spoken at Home) + 0.1924 (Urban) + 0.4095 (Public Insurance) + 0.6892 (Uninsured)	0.64807499 4764091
South American	**Logit(P)** = -3.0538 + 0.5964 (White) + 1.0823 (Other Language Spoken at Home) – 0.2789 (Age 0 to 21) – 0.315 (Age 22 to 40) + 0.5924 (Age 65+) -1.1961 (Public Insurance) – 1.0817 (Uninsured) + 0.5339 (Urban)	0.09634956 50503002
Spanish American	**Logit(P)** = 0.0163 + 1.0942 (American Indian or Alaska Native) + 0.0634 (White) – 1.2666 (Other Language Spoken at Home) – 0.6031 (Age 0 to 21) – 1.0966 (Age 22 to 40) + 0.2486 (Age 65+) – 0.3929 (Urban) – 0.5439 (Public Insurance) – 0.6745 (Uninsured)	0.25343703 9786678

*Age was kept as continuous. Source: 2021 Colorado Health Access Survey

Table 14. Logistic Regression Equations for CHAS Models Developed for the CO APCD

Subidentity Model	Coefficients and Logistic Regression Equation	Calculated Threshold
Caribbean/Central American	**Logit(P) =** -2.4978 – 0.3675 (American Indian or Alaska Native) + 0.0799 (Black or African American) + 0.5285 (Age 0 to 21) + 0.7576 (Age 22 to 40) + 0.0184 (Age 65+) – 0.00512 (Male) + 0.0319 (Public Insurance) – 0.4816 (Urban)	0.09070768 34337195
Chicano	**Logit(P) =** -1.0568 + 0.4998 (American Indian or Alaska Native) – 0.6181 (White) + 0.6186 (Some Other Race) – 0.0293 (Male) – 0.2368 (Age 0 to 21) + 0.127 (Age 22 to 40) – 0.4074 (Age 65+) + 0.5983 (Public Insurance) + 0.0747 (Urban)	0.27419955 4241577
Latinx*	**Logit(P) =** -2.1633 - 0.00406 (Age Continuous) + 0.5471 (American Indian or Alaska Native) + 0.3579 (Male) – 0.3325 (Public Insurance)	0.11415770 6227803
Mexican/Mexican American	**Logit(P) =** 0.3643 – 0.5326 (American Indian or Alaska Native) + 0.0882 (Black or African American) – 0.2032 (White) + 0.5951 (Age 0 to 21) + 0.5746 (Age 22 to 40) – 0.4347 (Age 65+) – 0.026 (Urban) + 0.4568 (Public Insurance)	0.67603300 0824139
South American	**Logit(P) =** -2.1648 + 0.0183 (American Indian or Alaska Native) + 0.2027 (White) + 0.5097 (Urban) – 1.3214 (Public Insurance) – 0.4476 (Age 0 to 21) – 0.4574 (Age 22 to 40) + 0.204 (Age 65+)	0.10465273 6000022
Spanish-American	**Logit(P) =** -0.6737 + 1.268 (American Indian or Alaska Native) + 0.4551 (White) – 0.4268 (Age 0 to 21) – 0.6692 (Age 22 to 40) + 0.6878 (Age 65+) – 0.3128 (Public Insurance) – 0.4019 (Urban)	0.25973481 4440422

*Age in the Latinx model was kept as continuous. Source: 2021 Colorado Health Access Survey

Lessons Learned

Plan for additional time invested to investigate covariates included in the model. Understanding the influence of each covariate on the model will take extra time. Some covariates, especially for specific populations, will not always be uniformly distributed. This will require reevaluation on how this impacts the model. Recategorization of the covariates will require harmonization across all datasets for comparability.

Investigate outside datasets that might contain the outcome of interest in the research. Understanding how models perform in real world settings can help refine them and aid researchers in understanding more about the relationships between the covariates and the outcome of interest. Take time to understand what other data systems might be collecting data of interest. This can help illuminate limitations and successes in the investigation.

Understand the population of interest. Different populations will have differing distributions and investigating these distributions will help identify covariates of interest to investigate in the analysis approach. These differences in distribution can have impacts on model performance.

What If Analysis

What if there were more data available on patient populations and those we are trying to understand in our communities? CHI investigated four scenarios to further explore covariates of interest that could increase the predictive power of the models. This exercise was designed to showcase the importance of data collection within existing or new databases that can be used for diagnosis or research purposes. CHI investigated four covariates: education, income, family size, and housing instability. These variables were chosen because they are socioeconomic variables that are related to other demographic characteristics investigated. Literature has shown that socioeconomic variables and racial/ethnic identity are intertwined concepts.[16,17,18]

Our analysis was completed using the CHAS training dataset used for the CHORDS models, which includes language as a covariate. The four covariates were all tested independently. How CHI coded these covariates is displayed in Table 15.

Table 15. Coding for Additional Covariates in What If Analysis and Reference Group

Covariate Name	Codes	Reference Group
Education	**1** = High school graduate or less, **2** = Some college or associate degree, **3** = College graduate or higher	**1**

Income*	**1** = 0 to 138% FPL, **2** = 139 to 400% FPL, **3** = Over 400% FPL	**2**
Family Size	**1** = 1 person, **2** = 2 people, **3** = More than 2 people	**2**
Housing Instability**	**1 =** Yes, experiencing unstable housing, **0** = No, not experiencing unstable housing	**0**

*FPL = Federal Poverty Level. **Housing instability was defined as those who reported they did not have stable housing in the next two months.

Table 16 includes the changes in the c statistic values with addition of each individual covariate in the CHORDS subidentity models. The models that had increased model fitness with addition of the covariate are highlighted in green.

Table 16. Pre- and Post-Adjustment of Each Covariate Compared to the Original Area Under the Curve (AUC) Estimate in the CHAS Models Built for the CHORDS

Subidentity Model	Pre-Adjusted C Statistic	Adjusted for Education C Statistic	Adjusted for Income C Statistic	Adjusted for Family Size C Statistic	Adjusted for Housing C Statistic
Caribbean/Central American	0.70	0.70	0.70	**0.71**	0.69
Chicano	0.69	**0.70**	0.69	**0.70**	0.69
Latinx	0.69	**0.76**	0.69	**0.70**	0.69
Mexican/Mexican American	0.61	**0.63**	0.61	**0.62**	**0.62**
South American	0.77	**0.80**	0.77	0.77	**0.78**
Spanish American	0.74	**0.75**	**0.75**	**0.75**	0.74

Source: 2021 Colorado Health Access Survey

The addition of income did not increase the predictive power of the CHAS models built for the CHORDS, which was not surprising given a correlation between insurance coverage and income. Instead, the inclusion of education or family size increased the predictive power of more of the models for the demographic variables. Housing, too, had more of an impact, although income does play a part in unstable housing as an affordability issue. As income was only broken into three categories, it may require additional analysis to understand the number of categories necessary to detect a statistical difference between the income groups, as 139% to 400% of FPL is a rather large income band. However, due to sample size, increasing the number of categories may just lead to discordance, making the model fitness questionable.

The inclusion of education created the largest increase in the c statistic for both the Latinx and South American models. For the Latinx model, addition of education presented a 0.04 p value for the overall model effects. Similarly, the addition of education into the South American model gave a p value <0.01, showing that education is highly predictive of both of these outcomes. Education, then, is an important demographic characteristic to include in model development for these specific subidentities. Education can also be used as a proxy for income as well, as it has been researched that these two variables are correlated.[19]

Addition of one more covariate that provides additional information about an individual can have a big impact on the overall predictive power of the logistic regression model. For example, addition of socioeconomic variables like education increased the predictive power of the South American model. Many health systems are already collecting information like housing stability and access to healthy foods; gathering more socioeconomic factors can enhance research and increase understanding of the people who live and access care in communities.

Lesson Learned

Expanded collection of social or demographic characteristics of a care-seeking population can increase the ability of researchers to answer questions that are, as of now, difficult to pursue. Health care systems are collecting information on social determinants of health that impact a person's ability to access quality care. Going a step further to understand the socioeconomic factors, like education or income, that affect a person and their health outcomes or health behaviors may be the next step to data collection efforts. Collection of these data will make approaches, like the one outlined in this paper, more able to answer research questions.

Limitations and Additional Discussion

The approach explained in this report can be especially helpful and instructive for other states and organizations investigating data disaggregation methods. However, Colorado has a distinct ethnic/racial makeup that differs from other states, and not all states or regions have the same data ecosystem as Colorado — both in the network of providers participating in a regional electronic health record database and the all-payer claims database. Colorado may be in a unique position to conduct this kind of inquiry into data disaggregation. That said, these results and this investigation shows the importance of partnerships in creating helpful and informative data systems. This research underscores how necessary these databases are to similar analyses that shed light on how policies and programs impact various communities.

One major assumption is that individuals represented in the CHAS dataset are representative of those in the CHORDS and CO APCD datasets. Preliminary descriptive statistics show that the overall distribution of certain characteristics is not the same across all three data systems. Additionally, both CHORDS and CO APCD datasets are limited to those who engage with the health care system. There are also geographic limitations as well, as the CHORDS network does not cover the entire state. This can insert bias in the analysis as there are differences in those who do and do not seek care in Colorado.

People who identify as Hispanic or Latino, as well as other ethnic or racial groups, face distinct barriers to seeking care, such as language limitations or the lack of culturally responsive care.[20] So, members of the Hispanic or Latino communities who access care may be intrinsically different than the overall Hispanic or Latino population in Colorado.

The predictive modelling approach also has limitations. Discerning differences among populations is only as good as the data that the methods are based on. Issues around data collection and missing data will impact the ability to create models that have high reliability and predictive power. Additionally, models are based on a specific number of covariates that are generally collected across multiple data systems. However, with a limited set of variables, the models might not have sufficient information to recognize enough predictive patterns to make the models valuable.

An example of this the inclusion of language spoken at home as a covariate. The CO APCD models did not include language as a covariate because of the absence of the variable in the claims database. As language spoken at home has been statistically proven in our modelling analysis to have an impact on the effectiveness of models developed to disaggregating ethnic data, so is it equally important to be collected within these data systems.

Although there were limitations, CHI was positioned to leverage Colorado’s unique data environment and apply these innovative data techniques.

Community Engagement Strategy

CHI engaged community organizations during the Phase II to frame the modelling approach with a community-driven lens. The following organizations were a part of the community outreach work:

- The Latino Community Foundation of Colorado
- Servicios de La Raza
- University of Colorado Denver, New Directions in Politics and Public Policy
- El Comité de Longmont
- Benefits in Action

CHI created a recruitment list based on a community ambassador program organized through the Metro Denver Partnership for Health. CHI chose groups that focus on serving and providing outreach to the Hispanic or Latino community in Colorado, with an emphasis on behavioral health and health care access.

Once the recruitment list was created, CHI then set up key informant interviews. An informant interview guide (see Appendix D) ensured uniformity in the information we collected.

Of the 13 organizations we reached out to, CHI was able to meet and speak with eight individuals across five organizations. Based on interviews, CHI created a community outreach matrix to track major themes that emerged in our discussions. Among them:

- Identity is complex and generational.
- Culturally competent care has a big impact on behavioral health utilization among Hispanic or Latino individuals.
- Fear and mistrust affect access to care, public benefits, and insurance enrollment.
- How an individual accesses care is based on their cultural context.
- It is important to disaggregate data, so Colorado has more information about health outcomes.
- Disaggregated data can then be used as educational tools to affect these health outcomes.

One of the more interesting findings from this outreach was how often identity and culture came up in our conversations. Each person lives within a cultural context, and this context that impacts many aspects of their life, including how they interact with the health care system. Within the Hispanic or Latino community there is a complex ecosystem of other identities. In Colorado, one of the major ethnic identities within the Hispanic or Latino community is Mexican. This means the Mexican culture and dialect may often govern how Spanish is spoken, taught, and the kinds of social norms that are represented.

This major identity has an impact on other identities that aren't as common in the state. How does someone who identities as Caribbean, for example, feel represented in a Hispanic or Latino culture when they make up only 4% of this category? What language barriers are inadvertently put in place because of the primary Spanish dialect from Mexico? This and more can impact a person's ability to access Colorado's health care system.

Another important concept is how people within each subculture view behavioral health care. Some cultures depend on the family unit in dealing with behavioral health issues. Others take a different approach or may not identify with the way mental health or substance use is talked about in Colorado. Each individual lens and experience influence this discussion differently, so getting people the help they need is more complicated than just having language translation or interpretation services available.

All of these conversations affected our overall methodological approach. A person's preferred language was a key part of these discussions; indeed, the models showcased how important this factor is in predicting someone's subidentity. Continuing these conversations and building trust with community members and groups in Colorado is important to making data and information available to those who can use it to better serve their client base.

The community engagement work also shaped next steps for Phase III. The community groups identified behavioral health as an integral component of their work and said getting people access to needed care is a priority. As a result, identifying groups that might need additional outreach and engagement to increase their use of behavioral health care will be the foundation for the Phase III research approach.

All participants were provided a $50 gift card as well as a memo that included data from the 2021 CHAS on the disaggregated Hispanic or Latino identities. Participants also received a data workbook with additional information.

Lessons Learned

Community engagement strategies should be built into the research approach and empower community voice in the process. Research questions and analyses need to be built with the community partners, as these findings are meant to frame the work moving forward to better understand the communities that are underrepresented in the data. Giving these communities a voice and showing that they are seen is crucial for community investment in these strategies.

Engage the community to create a two-way learning experience. Qualitative information can be used to shape research questions and understand unmet community needs. This engagement can then create partnerships where all participants can benefit from the information, guide practices, and develop programs that are community informed.

Next Steps for Our Work

Using the information and models developed during Phase II, Phase III of the investigation will start at the end of 2022 and wrap up in June of 2023. This phase will focus on a use-case analysis of the models, investigating differences in behavioral health utilization and outcomes across the subidentities examined.

The main research questions that Phase III will seek to answer are:

- How do mental health diagnoses and utilization of mental health services differ between racial/ethnic groups in Colorado? Among people who identify as Hispanic/Latino?

- Are certain groups more likely to receive behavioral health screenings/services compared to others?
- Are there differences in the mode in which care is delivered between racial/ethnic groups (telehealth vs. in-person visits)?
- Are there differences in the types of providers delivering behavioral health care between racial/ethnic groups (physician, psychologist, nurse practitioner, etc.)?

The community groups working on this project with CHI have been notified of the next phase of work. CHI will keep them informed along the way and consult with them about the findings and their implications. CHI plans to produce a report on the Phase III results and may host a webinar to present the information to a wider audience.

Future work may also involve the 2023 CHAS, in which CHI plans to continue collecting disaggregated ethnicity data to understand the subidentities of Hispanic or Latino respondents. Using these data, CHI could investigate the methodology and refine the models, as combining the 2021 and 2023 CHAS datasets provides a larger sample size to work with for modelling development. A separate model for Caribbean and Central American subidentities could be created from this research.

CHI has completed preliminary analysis of the CHORDS and CO APCD datasets for Phase III. Those data can be provided upon request.

Conclusion

Disaggregated data on racial or ethnic characteristics can be useful in many ways. These data allow us to understand health outcomes that show up differently across subgroups and identify those communities within geographies or jurisdictions.

Disaggregated data also informs the work of organizations that serve these communities. Being equipped with data and information about their target populations can help these organizations seek additional resources, make decisions, expand their programming, and subsequently serve more people.

CHI presents a methodological approach to disaggregating data in this report. The research team embarked on this exploration to understand the extent to which collecting disaggregated ethnicity data could be leveraged to disaggregate other datasets. Of the six Hispanic/Latino subgroups studied, the models performed well in predicting three of the subgroups. Refinement and exploratory analyses of the remaining models will also take place to increase the predictive power of the models. A predictive modelling approach is easily reproducible by other organizations or entities. States with a higher percentage of people who identify as Hispanic or Latino can also benefit from this research and this approach.

Lessons taken from this approach include:

- Data disaggregation methods should be community informed to ensure that the collection, management, and analysis of racial/ethnic identity data are correctly interpreted, actionable, and useful to community organizations' work and missions.
- Local data expertise should be leveraged not only to increase efficiency of the development of the statistical methods but also to increase awareness and emphasize the importance of disaggregating data within more data systems.
- Data collection efforts set the stage for this work — invest the time and resources into available data systems to gather representative disaggregated data.

The more we understand about the people within these communities, the more we can work together with community groups, members, and leaders toward finding ways to ensure Coloradans can live healthy, successful lives.

Appendix A. 2021 CHAS Survey Questionnaire Disaggregated Hispanic or Latino Ethnicity Items

D1. (Are you/is TARGET) Hispanic or Latino?

1 Yes
2 No, not of Hispanic origin
8 Don’t know
9 Refused

(ASK if D1 = 1)
D1a. Please indicate how (you identify/TARGET identifies) or represent (yourself/themselves).

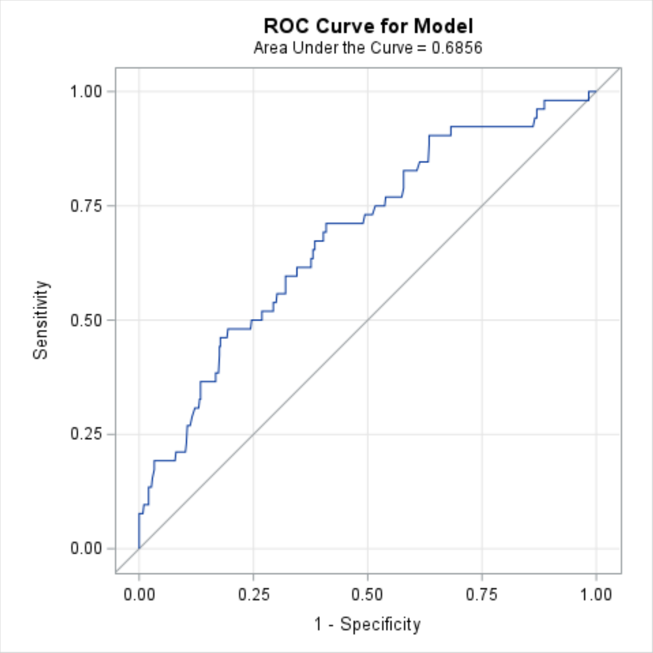
1 Yes
2 No
8 Don’t know
9 Refused

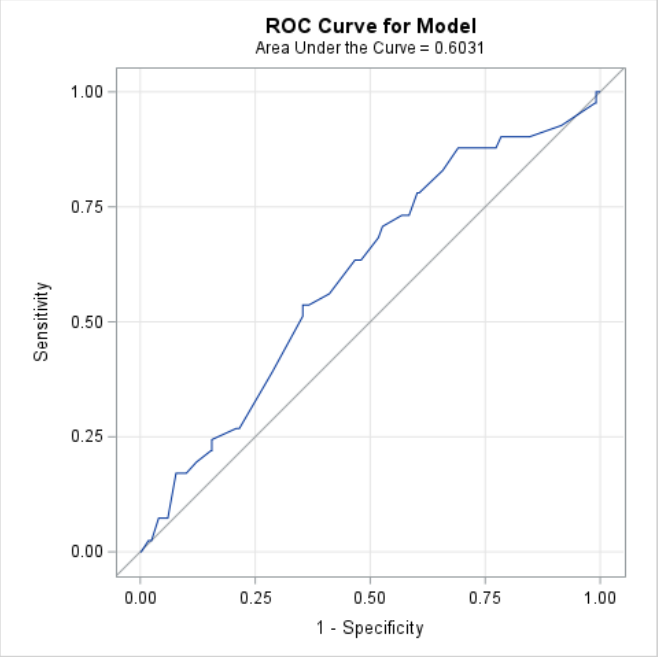
a. Mexican/Mexican American
b. Chicano
c. Central American (El Salvador, Guatemala, Honduras, Nicaragua, Panamá, etc.)
d. South American (Chile, Colombia, Ecuador, Perú, Venezuela, etc.)
e. Caribbean (Cuba, Dominican Republic)
f. Latinx
g. Spanish-American (from Spain)
h. Something else (Specify: ___________)

Note: Respondents may select more than one choice on question D1a.

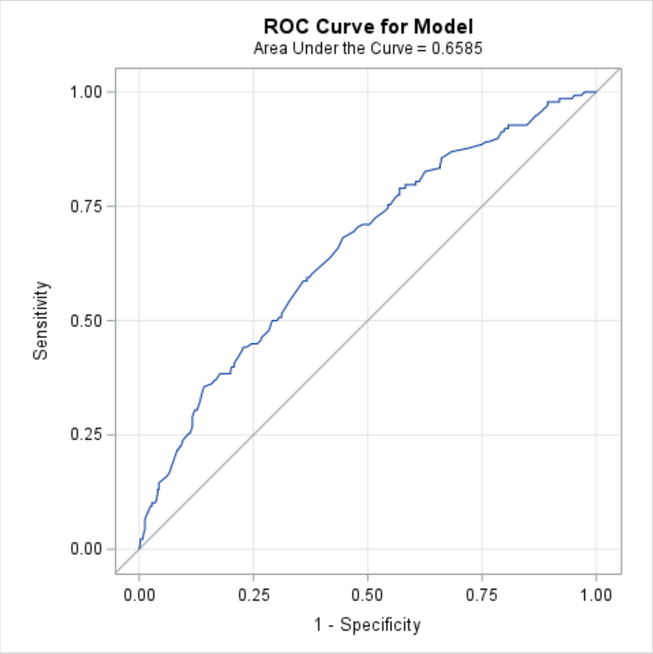
Appendix B. ROC Curve Output

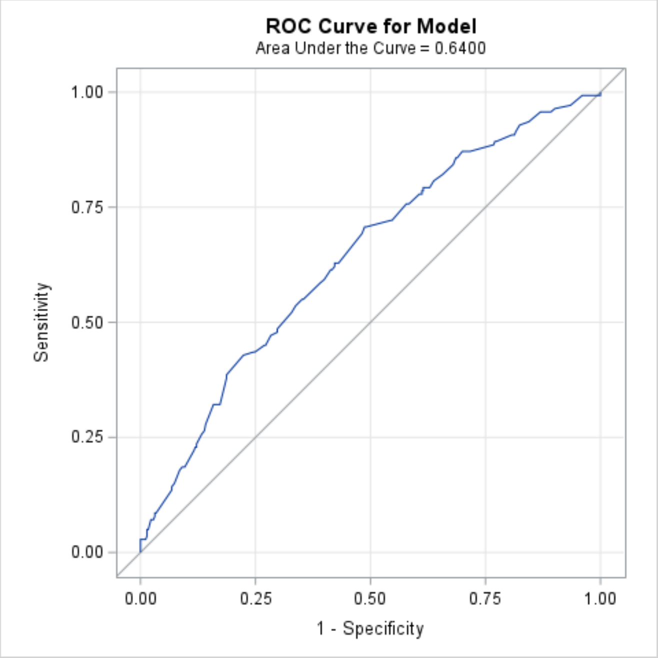
Subidentity Caribbean/Central American ROC Curve Output (CHORDS and CO APCD)



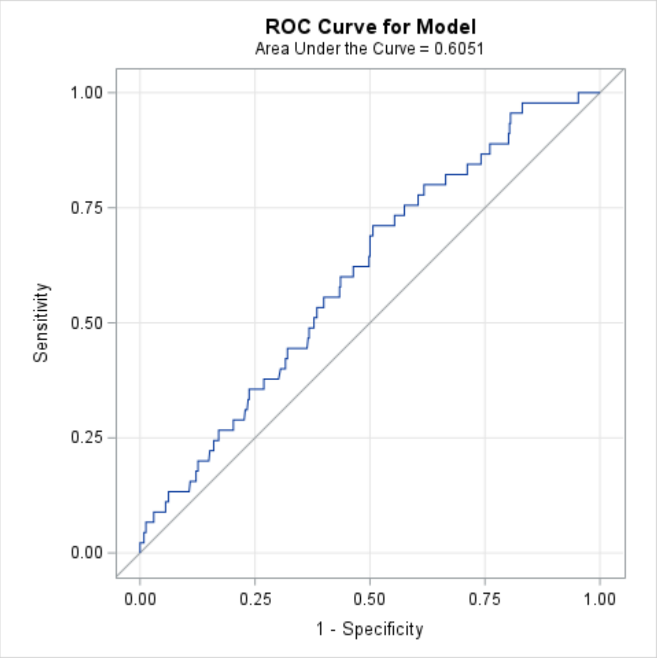


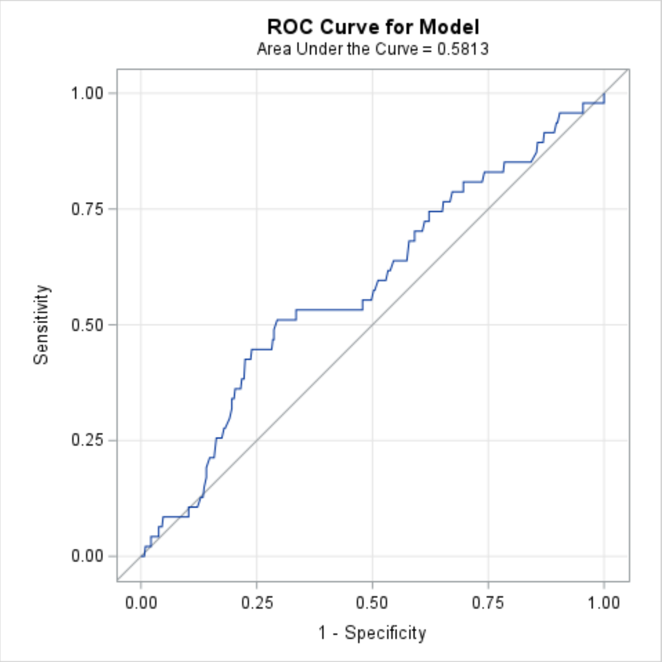
Subidentity Chicano ROC Curve Output (CHORDS and CO APCD)



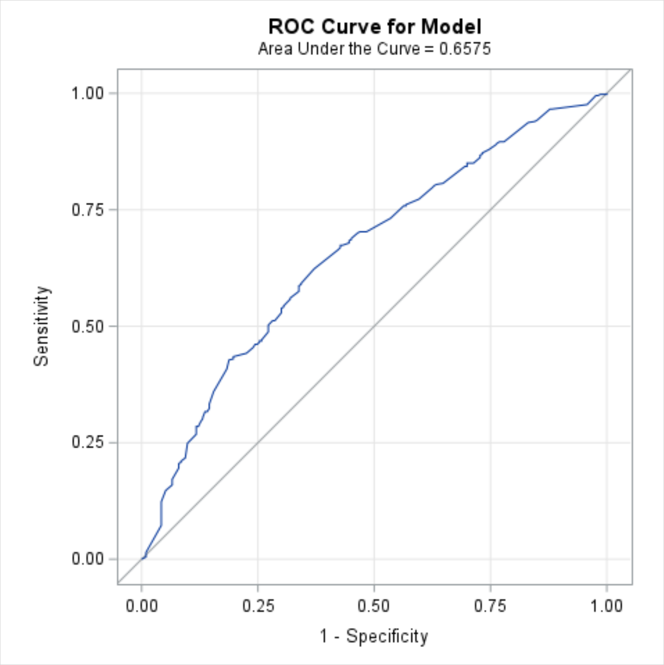


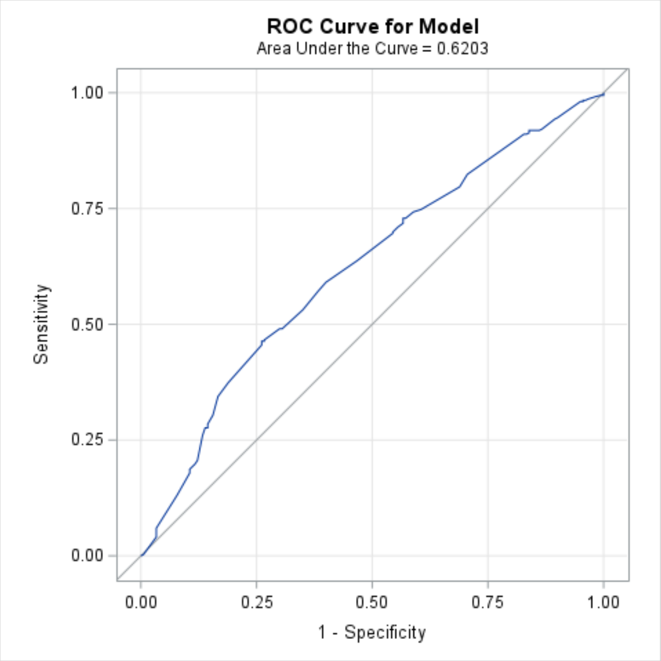
Subidentity Latinx ROC Curve Output (CHORDS and CO APCD)



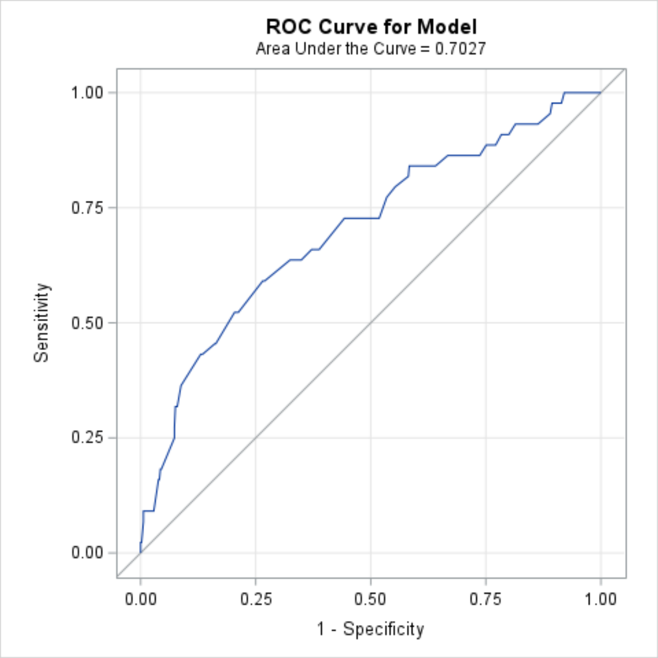


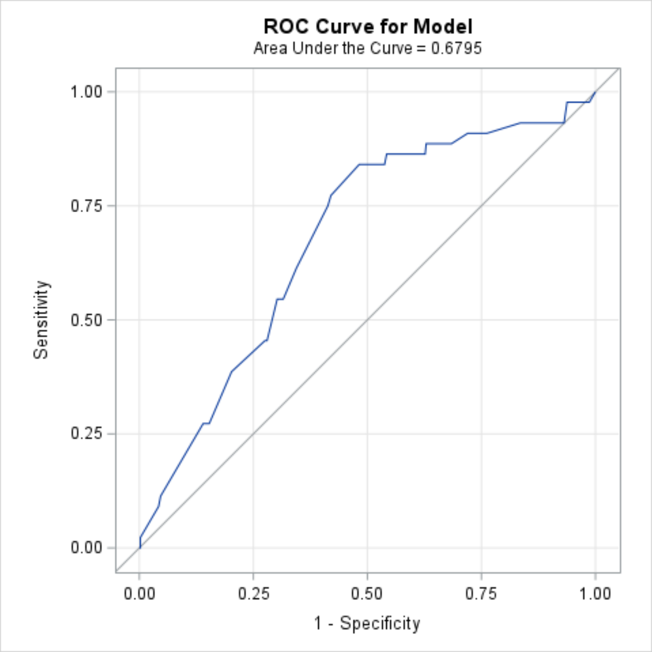
Subidentity Mexican/Mexican American ROC Curve Output (CHORDS and CO APCD)



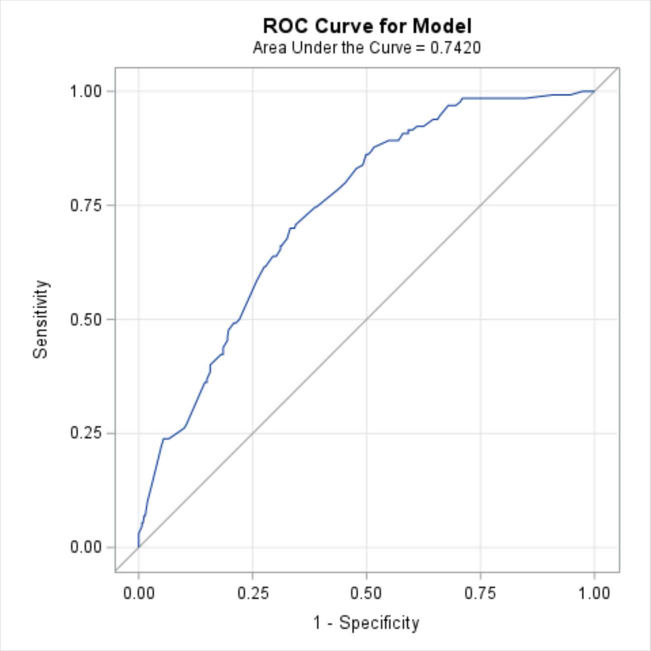


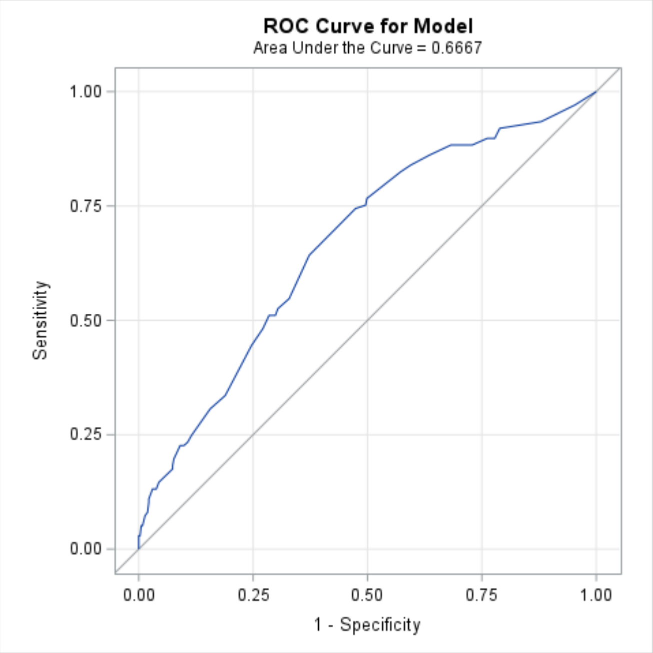
Subidentity South American ROC Curve Output (CHORDS and CO APCD)





Subidentity Spanish American ROC Curve Output (CHORDS and CO APCD)





Appendix C. Predictive Model Output from PROC LOGISTIC Procedure

CHAS Models Built for the CHORDS

Caribbean/Central American

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
American Indian or Alaska Native	1	0.3013	0.5831
White	1	0.0035	0.9530
Black or African American	1	5.7077	0.0169
Other Language Spoken at Home	1	13.4134	0.0002
Age (Continuous)	1	2.8313	0.0924
Insurance Coverage Type	2	1.5391	0.4632

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.3724	0.4629	26.2645	<.0001
American Indian or Alaska Native	1	1	-0.4269	0.7778	0.3013	0.5831
White	1	1	0.0207	0.3515	0.0035	0.9530
Black or African American	1	1	1.5247	0.6382	5.7077	0.0169
Other Language Spoken at Home	1	1	1.2600	0.3440	13.4134	0.0002
Age (Continuous)		1	-0.0141	0.00841	2.8313	0.0924
Insurance Coverage Type	2	1	-0.4137	0.3554	1.3555	0.2443
Insurance Coverage Type	3	1	-0.3664	0.4912	0.5562	0.4558

Chicano

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
American Indian or Alaska Native	1	0.0176	0.8945
White	1	6.2462	0.0124
Age (Grouped)	3	1.9974	0.5729
Other Language Spoken at Home	1	5.9263	0.0149

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Gender	1	0.0038	0.9506
Rurality	1	4.3156	0.0378
Insurance Coverage Type	2	11.3986	0.0033

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.5710	0.2734	4.3606	0.0368
American Indian or Alaska Native	**1**	1	0.0487	0.3668	0.0176	0.8945
White	**1**	1	-0.6174	0.2470	6.2462	0.0124
Age Group	**1**	1	-0.2303	0.2755	0.6992	0.4031
Age Group	**2**	1	0.0884	0.2600	0.1155	0.7339
Age Group	**4**	1	0.2964	0.4479	0.4380	0.5081
Other Language Spoken at Home	**1**	1	-0.5499	0.2259	5.9263	0.0149
Gender	**1**	1	0.0130	0.2096	0.0038	0.9506
Rurality	**1**	1	-0.4460	0.2147	4.3156	0.0378
Insurance Coverage Type	**2**	1	0.6793	0.2305	8.6822	0.0032
Insurance Coverage Type	**3**	1	-0.3197	0.3901	0.6714	0.4126

Latinx

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
American Indian or Alaska Native	1	1.6228	0.2027
White	1	0.0335	0.8549
Age (Continuous)	1	0.1397	0.7086
Gender	1	0.6297	0.4275
Other Language Spoken at Home	1	0.4085	0.5227
Insurance Coverage Type	2	0.1082	0.9473
Rurality	1	3.8815	0.0488

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.9072	0.5940	23.9571	<.0001
American Indian or Alaska Native	**1**	1	0.6688	0.5250	1.6228	0.2027
White	**1**	1	0.0660	0.3611	0.0335	0.8549
Age (Continuous)		1	-0.00310	0.00829	0.1397	0.7086
Gender	**1**	1	-0.2521	0.3177	0.6297	0.4275
Other Language Spoken at Home	**1**	1	0.2186	0.3421	0.4085	0.5227
Insurance Coverage Type	**2**	1	0.1141	0.3520	0.1050	0.7459
Insurance Coverage Type	**3**	1	0.0791	0.5318	0.0222	0.8817
Rurality	**1**	1	0.7680	0.3898	3.8815	0.0488

Mexican/Mexican American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	0.5480	0.4591
Black or African American	1	1.5987	0.2061
White	1	0.1605	0.6887
Age (Grouped)	3	10.3803	0.0156
Other Language Spoken at Home	1	11.2122	0.0008
Rurality	1	1.0840	0.2978
Insurance Coverage Type	2	7.3648	0.0252

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.2228	0.2219	1.0089	0.3152
American Indian or Alaska Native	**1**	1	-0.2429	0.3281	0.5480	0.4591
Black or African American	**1**	1	-0.6040	0.4777	1.5987	0.2061
White	**1**	1	-0.0784	0.1956	0.1605	0.6887
Age Group	**1**	1	0.7151	0.2276	9.8691	0.0017

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Age Group	**2**	1	0.4056	0.2188	3.4368	0.0638
Age Group	**4**	1	0.5501	0.4538	1.4698	0.2254
Other Language Spoken at Home	**1**	1	0.6410	0.1914	11.2122	0.0008
Rurality	**1**	1	0.1924	0.1848	1.0840	0.2978
Insurance Coverage Type	**2**	1	0.4095	0.2054	3.9766	0.0461
Insurance Coverage Type	**3**	1	0.6892	0.3083	4.9976	0.0254

South American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
White	1	2.7253	0.0988
Other Language Spoken at Home	1	9.5341	0.0020
Age (Grouped)	3	1.9792	0.5767
Insurance Coverage Type	2	8.5054	0.0142
Rurality	1	2.0031	0.1570

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-3.0538	0.4500	46.0607	<.0001
White	**1**	1	0.5964	0.3613	2.7253	0.0988
Other Language Spoken at Home	**1**	1	1.0823	0.3505	9.5341	0.0020
Age Group	**1**	1	-0.2789	0.4328	0.4154	0.5193
Age Group	**2**	1	-0.3150	0.4129	0.5819	0.4456
Age Group	**4**	1	0.5924	0.7325	0.6540	0.4187
Insurance Coverage Type	**2**	1	-1.1961	0.4614	6.7197	0.0095
Insurance Coverage Type	**3**	1	-1.0817	0.6405	2.8519	0.0913
Rurality	**1**	1	0.5339	0.3773	2.0031	0.1570

Spanish American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	9.6958	0.0018
White	1	0.0703	0.7909
Other Language Spoken at Home	1	23.6934	<.0001
Age (Grouped)	3	16.6010	0.0009
Rurality	1	3.0659	0.0800
Insurance Coverage Type	2	5.7178	0.0573

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.0163	0.2510	0.0042	0.9483
American Indian or Alaska Native	**1**	1	1.0942	0.3514	9.6958	0.0018
White	**1**	1	0.0634	0.2392	0.0703	0.7909
Other Language Spoken at Home	**1**	1	-1.2666	0.2602	23.6934	<.0001
Age Group	**1**	1	-0.6031	0.2731	4.8774	0.0272
Age Group	**2**	1	-1.0966	0.2949	13.8275	0.0002
Age Group	**4**	1	0.2486	0.4755	0.2734	0.6011
Rurality	**1**	1	-0.3929	0.2244	3.0659	0.0800
Insurance Coverage Type	**2**	1	-0.5439	0.2629	4.2784	0.0386
Insurance Coverage Type	**3**	1	-0.6745	0.4460	2.2876	0.1304

CHAS Models Built for the CO APCD

Caribbean/Central American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	0.2373	0.6262
Black or African American	1	0.0055	0.9408
Age (Grouped)	3	3.4428	0.3283

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Gender	1	0.0002	0.9878
Insurance Coverage Type	1	0.0082	0.9278
Rurality	1	1.9932	0.1580

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.4978	0.4234	34.8016	<.0001
American Indian or Alaska Native	**1**	1	-0.3675	0.7545	0.2373	0.6262
Black or African American	**1**	1	0.0799	1.0751	0.0055	0.9408
Age Group	**1**	1	0.5285	0.4428	1.4245	0.2327
Age Group	**2**	1	0.7576	0.4345	3.0404	0.0812
Age Group	**4**	1	0.0184	0.8280	0.0005	0.9823
Gender	**1**	1	-0.00512	0.3342	0.0002	0.9878
Insurance Coverage Type	**2**	1	0.0319	0.3516	0.0082	0.9278
Rurality	**1**	1	-0.4816	0.3411	1.9932	0.1580

Chicano

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	1.9082	0.1672
White	1	6.6755	0.0098
Some Other Race	1	2.1703	0.1407
Gender	1	0.0196	0.8887
Age (Grouped)	3	2.3818	0.4970
Insurance Coverage Type	1	7.1396	0.0075
Rurality	1	0.1125	0.7374

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0568	0.2685	15.4906	<.0001
American Indian or Alaska Native	**1**	1	0.4998	0.3618	1.9082	0.1672
White	**1**	1	-0.6181	0.2392	6.6755	0.0098
Some Other Race	**1**	1	0.6186	0.4199	2.1703	0.1407
Gender	**1**	1	-0.0293	0.2096	0.0196	0.8887
Age Group	**1**	1	-0.2368	0.2747	0.7429	0.3887
Age Group	**2**	1	0.1270	0.2608	0.2372	0.6262
Age Group	**4**	1	-0.4074	0.4359	0.8735	0.3500
Insurance Coverage Type	**2**	1	0.5983	0.2239	7.1396	0.0075
Rurality	**1**	1	0.0747	0.2227	0.1125	0.7374

Latinx

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age (Continuous)	1	0.2778	0.5981
American Indian or Alaska Native	1	1.1267	0.2885
Gender	1	1.2866	0.2567
Insurance Coverage Type	1	1.0449	0.3067

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1633	0.3720	33.8134	<.0001
Age (Continuous)		1	-0.00406	0.00770	0.2778	0.5981
American Indian or Alaska Native	**1**	1	0.5471	0.5154	1.1267	0.2885
Gender	**0**	1	0.3579	0.3155	1.2866	0.2567
Insurance Coverage Type	**2**	1	-0.3325	0.3252	1.0449	0.3067

Mexican/Mexican American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	2.3429	0.1259
Black or African American	1	0.0204	0.8864
White	1	1.0343	0.3092
Age (Grouped)	3	12.4369	0.0060
Rurality	1	0.0174	0.8951
Insurance Coverage Type	1	4.7614	0.0291

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.3643	0.2121	2.9494	0.0859
American Indian or Alaska Native	**1**	1	-0.5326	0.3480	2.3429	0.1259
Black or African American	**1**	1	0.0882	0.6175	0.0204	0.8864
White	**1**	1	-0.2032	0.1998	1.0343	0.3092
Age Group	**1**	1	0.5951	0.2428	6.0080	0.0142
Age Group	**2**	1	0.5746	0.2367	5.8913	0.0152
Age Group	**4**	1	-0.4347	0.4096	1.1265	0.2885
Rurality	**1**	1	-0.0260	0.1969	0.0174	0.8951
Insurance Coverage Type	**2**	1	0.4568	0.2093	4.7614	0.0291

South American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	0.0008	0.9772
White	1	0.3582	0.5495
Rurality	1	1.8686	0.1716
Insurance Coverage Type	1	8.0366	0.0046
Age (Grouped)	3	2.0646	0.5591

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-2.1648	0.3774	32.9107	<.0001
American Indian or Alaska Native	**1**	1	0.0183	0.6380	0.0008	0.9772
White	**1**	1	0.2027	0.3388	0.3582	0.5495
Rurality	**1**	1	0.5097	0.3729	1.8686	0.1716
Insurance Coverage Type	**2**	1	-1.3214	0.4661	8.0366	0.0046
Age Group	**1**	1	-0.4476	0.4282	1.0927	0.2959
Age Group	**2**	1	-0.4574	0.3992	1.3128	0.2519
Age Group	**4**	1	0.2040	0.8370	0.0594	0.8074

Spanish American

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
American Indian or Alaska Native	1	11.9645	0.0005
White	1	4.2077	0.0402
Age (Grouped)	3	12.6815	0.0054
Insurance Coverage Type	1	1.6760	0.1955
Rurality	1	3.4630	0.0628

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.6737	0.2267	8.8282	0.0030
American Indian or Alaska Native	**1**	1	1.2680	0.3666	11.9645	0.0005
White	**1**	1	0.4551	0.2218	4.2077	0.0402
Age Group	**1**	1	-0.4268	0.2684	2.5298	0.1117
Age Group	**2**	1	-0.6692	0.2808	5.6803	0.0172
Age Group	**4**	1	0.6878	0.4098	2.8168	0.0933
Insurance Coverage Type	**2**	1	-0.3128	0.2416	1.6760	0.1955
Rurality	**1**	1	-0.4019	0.2160	3.4630	0.0628

Appendix D. Community Engagement Interview Guide

Introduction

Our main aims for speaking to community are the following:

1. Consult with community on the importance and usefulness of disaggregated data for Hispanic/Latino groups and explore gaps in current CHAS data that would help inform and advance community work if data were available.
2. Present indicators of behavioral health by disaggregated Hispanic/Latino groups from the CHAS in order to collect insight on the meaning and usefulness of this data among Hispanic/Latino groups.
3. Consult on the social and cultural barriers to behavioral health care within Hispanic/Latino communities to help inform and identify future metrics.

You have been contacted because you are a prominent leader within an organization that serves Hispanic/Latino/a communities here in Colorado.

Part 1 Questions

Present available CHAS metrics on more granular (disaggregated) identities within the Hispanic/Latino community groups and leaders.

- *Mental health status*
- *Uninsured*

1. Are there any reactions to this data that you would like to share with us?
2. From your perspective, are we missing anything here that would be helpful to the communities you work with and serve?

Part 2 Questions

1. How would access to disaggregated data be helpful to your work, if at all?
    a. In what ways do you prefer to access data?
2. Can you tell us about a time where the absence of disaggregated data impacted your work directly?
3. What type of metrics and outcomes would be most useful/actionable to see at the disaggregated level?

Part 3 Questions

1. To what extent do people’s experiences with certain systems (such as health care for example) differ by their specific Hispanic or Latinx ethnic identity or country of origin?

    a. To what extent do you see differences in the types of barriers groups experience by disaggregated group?
2. How can this data help address any specific barriers?
3. Is there anyone that you think we should talk to?

Closing and follow-up

1. How can CHI help you and/or your organization?

Anything else we should have touched on during this call that you think is important?

[1] Simón, Y. (2020). Latino, Hispanic, Latinx, Chicano: The History Behind the Terms. History. https://www.history.com/news/hispanic-latino-latinx-chicano-background
[2] Colorado Health Institute. (2022). The Colorado Health Access Survey. https://www.coloradohealthinstitute.org/research/colorado-health-access-survey-chas
[3] Center for Improving Value in Health Care. (2022). CO APCD Info. https://www.civhc.org/get-data/co-apcd-info/
[4] Colorado Health Observation Regional Data Service. (2022). CHORDS. https://www.coloradohealthinstitute.org/research/CHORDS
[5] Bursac, Z., et al. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine. 3: 17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2633005/
[6] Austin, P. and Steyerberg, E. (2021). Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Medical Research Methodology. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-82
[7] Fagerland, M and Hosmer. H. (2012). A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models. The Stata Journal. 3: 447-453.
[8] Kramer, A, and Zimmerman, J. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Critical Care Medicine. 35(9): 2056-6.
[9] JHajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian Journal of Internal Medicine. 4(2): 627-635.
[10] Zou, K, O'Malley, J, and Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation. 115:654-657.
[11] Pandey, M and Jain, A. (2016). ROC Curve: Making way for correct diagnosis. PharmaSUG. https://www.pharmasug.org/proceedings/2016/SP/PharmaSUG-2016-SP11.pdf
[12] Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013). Applied Logistic Regression. P 177. John Wiley and Sons. DOI:10.1002/9781118548387.
[13] Ogundimu, E., Altman, D., and Collins, G. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. 76: 175-182.
[14] Colorado Health Institute. (2022). Language and Culturally Responsive Care in Colorado: Barriers, Access, and Room for Improvement. https://www.coloradohealthinstitute.org/research/language-and-culturally-responsive-care-colorado
[15] Colorado Health Institute. (2022). Diverse State, Diverse Needs: Coloradans' Needs and Experiences Highlight Demand for Culturally Responsive Care. https://www.coloradohealthinstitute.org/research/culturally-responsive-care-Colorado
[16] American Psychological Association. (2017). Ethnic and Racial Minorities & Socioeconomic Status. https://www.apa.org/pi/ses/resources/publications/minorities#:~:text=The%20relationship%20between%20SES%2C%20race,SES%2C%20race%2C%20and%20ethnicity
[17] Anderson, N. (2004). 9, Race/Ethnicity, Socioeconomic Status, and Health. Washington (DC): National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK25526

[18] Noël, R. (2018). Race, Economics, And Social Status. U.S. Bureau of Labor Statistics. https://www.bls.gov/spotlight/2018/race-economics-and-social-status/pdf/race-economics-and-social-status.pdf

[19] U.S. Bureau of Labor Statistics. (2021). Education pays. https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

[20] Colorado Health Institute. (2022). Language and Culturally Responsive Care in Colorado: Barriers, Access, and Room for Improvement. https://www.coloradohealthinstitute.org/research/language-and-culturally-responsive-care-colorado