Colorado Ethnicity Data Disaggregation Feasibility Assessment

The Case for Working Across Datasets

June 18, 2021

COLORADO
HEALTH
INSTITUTE

Informing Strategy. Advancing Health.

Executive Summary

More than a fifth of Coloradans (21.8%) identified as Hispanic/Latinx in 2019.[1] However, data on more specific ethnic identities within this group are limited. This reduces the understanding of health outcomes, access to care, and use of care given that these areas might differ among more specific race/ethnic groups in Colorado.

Data disaggregation is used to uncover populations often hidden in the data. It describes the process of collecting and analyzing information on granular sub-categories of racial/ethnic identity, a process that can reveal disparities that aggregated data cannot. To bridge gaps in data reporting, the Colorado Health Institute (CHI), in partnership with the UCLA Center for Health Policy Research and the Robert Wood Johnson Foundation, explored a strategy of disaggregating racial/ethnic health data in Colorado. CHI is collecting disaggregated data on ethnicity and assessed the feasibility of using disaggregated data across three sources: surveys, insurance claims, and electronic medical records.

Working across these systems could provide information on health access and health outcomes for people identifying with specific ethnic groups within the Hispanic/Latinx community in Colorado. Based on a review of the literature, CHI developed a toolbox of approaches for achieving this data disaggregation. The statistical methods that CHI included in the toolbox are categorized as either "linking" or "expansion" strategies.

CHI outlined a multistage Theory of Change model for addressing a barrier and how data disaggregation may ultimately inform policy changes that improve health.

In accordance with the Theory of Change, CHI developed and applied criteria to disaggregate data in Colorado datasets. Two approaches — race bridging and predictive modeling — were identified as promising methods for disaggregating data among Coloradans who identify as Hispanic/Latinx. Race bridging is a method employed by the National Center for Health Statistics, the U.S. Census Bureau, and other agencies to create estimates that can be trended over time when racial/ethnic categorizations change. Because CHI updated the race/ethnicity categories on the Colorado Health Access Survey (CHAS) in 2021, this technique will be integral in creating comparable estimates across all survey years of data.

In combination with race bridging, CHI will use a predictive modelling approach to expand current data systems with disaggregated data. To achieve this, CHI will model the common variables across all datasets to understand their association with disaggregated subgroups in the CHAS survey. CHI will then apply this model to both the Colorado Health Observation Regional Data Service (CHORDS) and the Colorado All-Payer Claims Database (CO APCD) to investigate whether it is possible to predict the likelihood that an individual identifies as a specific disaggregated subgroup.

Engaging stakeholders is a key ingredient in data disaggregation efforts. CHI learned through the feasibility assessment that engaging potential data users, stewards (those who administer the data), and communities most affected by the findings (some of which also may be users and stewards) is important in the analysis and dissemination of disaggregated data.

In this report, CHI investigated possible strategies to disaggregate data and important considerations when pursuing these techniques. Using these strategies, CHI investigated the feasibility of methods to disaggregate data on Hispanic/Latinx ethnicity across Colorado's own data systems. CHI has proposed a Colorado use case and analysis plan for accomplishing data disaggregation, including the process for sharing the results of this analysis and important lessons learned from the assessment. These next steps will be integral in continuing the efforts to obtain disaggregated data for more specific identities in Colorado's communities.

Acknowledgements

Colorado Health Institute Team
Lindsey Whittington, Lead Author
Kristi Arellano
Jeff Bontrager
Brian Clark
Deanna Geldens
Emily Johnson

Collaborators
Emily Bacon, PhD, Bacon Analytics LLC
Thomas Belin, PhD, UCLA Department of Biostatistics
Anna Furniss, MS, University of Colorado Anschutz Medical Campus
Maria de Jesus Diaz-Perez, PhD, Center for Improving Value in Health Care
Rachel Zucker, MPH, PMP, University of Colorado Anschutz Medical Campus

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

Introduction: Why Disaggregate Data?

Disparities in health outcomes, access to care, utilization, and quality of care between people of different races/ethnicities have been well documented. To understand these disparities — as well as strengths and areas where groups excel — data with accurate and robust reporting must be available to inform policy initiatives and other programmatic changes to address these inequities. Unfortunately, data quality and availability are often limited, making it hard to characterize existing differences or similarities.

In general, data sources used to quantify these disparities rely on aggregated racial categories: African American/Black, Asian American/Native Hawaiian/Pacific Islander, American Indian/Alaska Native, and white. Many data sources also include a separate indicator of Hispanic/Latinx ethnicity, which is often combined with the racial categories to create mutually exclusive groups such as non-Hispanic/Latinx African American/Black.

While some data sources use more categories, others must use fewer because of small population sizes and other reporting issues. Because these categories are so broad, they tend to mask differences within these groups. Understanding and addressing health disparities requires greater granularity within the data.

An example of the gap in reporting came during the preparation for the 2020 census, when the U.S. Census Bureau considered adding a Middle Eastern or North African response option to the current classification. In 2010, the Census Bureau found that many people who identify as Middle Eastern or North African did not know how to respond and felt excluded from the existing categories.[2] Efforts to add this category stalled in 2018, although efforts may be reconsidered with hopes of improving data collection to enhance available demographic information on the census.[3]

Gathering more specific race/ethnicity data — referred to as data disaggregation — provides visibility to groups that might otherwise be invisible in current estimates of health outcomes. Disaggregating data on race/ethnicity is a collaborative process between those who collect the data and those who are represented in the numbers. By making more specific race/ethnicity data available, communities can inform policy to address existing disparities and highlight strengths that might be masked by current methods.

To address reporting issues at the local level, the Colorado Health Institute (CHI) explored a strategy of disaggregating health data in Colorado. CHI focused on key data systems that were rich sources of information about health outcomes and health care access. These data systems included the Colorado Health Access Survey (CHAS), a biennial survey of health data administered by CHI; the Colorado Health Observation Regional Data Service (CHORDS), a regional network of health systems and providers that bring together their electronic health records for public health research; and the Colorado All-Payer Claims Database (CO APCD), housed within the Center for Improving Value in Health Care (CIVHC). Specifically, this report assesses the feasibility of methods to disaggregate data on Hispanic/Latinx ethnicity across Colorado's data systems. It also identifies potential applications of disaggregated data, recommendations for next steps in the process, and lessons learned.

Guiding Questions

To understand the feasibility of linking or expanding information across data systems to disaggregate race/ethnicity data in Colorado, CHI developed a set of guiding questions to inform the feasibility assessment. Those questions included:

- To what degree is data disaggregation achievable by linking or expanding datasets?
- What new research questions are answerable by linking or expanding datasets?
- What linking or expansion approaches are feasible?
- To what extent are there disclosure or confidentiality risks?
- What concerns do health care consumers have?
- What factors must be considered when linking or expanding available data in service to data disaggregation?

A Multi-Stage Approach for Data Disaggregation

To address these questions, Figure 1 displays the multistage conceptual diagram — called a Theory of Change — that identifies how disaggregating data on Hispanic/Latinx ethnicity will ultimately contribute to improved health. CHI used the Theory of Change to guide this feasibility study.

The first step in the Theory of Change identifies the specific problem or challenge — smaller ethnic groups going unrepresented or unidentified in available data — as well as the guiding questions that address this problem, listed in the previous section.

This feasibility study represents the culmination of Step 2: Exploration of Methods and Data. Step 2 began with identifying available data in Colorado and engaging the respective organizations or entities that manage those data. These sources included the CHAS, CHORDS, and CO APCD.

CHI then conducted a literature review and engaged with statisticians to identify possible methods for disaggregating data. CHI then developed and applied criteria to select the most appropriate methods given Colorado-specific considerations and constraints. Finally, CHI developed a recommended analysis plan for analyzing the data. This report is generally structured to reflect the components of Step 2.

Subsequent steps in the Theory of Change involve obtaining the data, applying the analysis plan, and sharing the results (steps 3 and 4). CHI anticipates that the results will be used to inform policy and practice (Step 5), address disparities, and build on strengths within Colorado's Hispanic/Latinx communities (Step 6), and ultimately improve health and well-being (Step 7).

CHI acknowledges that there are many other factors beyond policy and practice that determine a person's health, such as environmental factors and socioeconomic status. This Theory of Change, however, focuses on how the disaggregation of racial and ethnic information specifically could lead to improved outcomes.

An important principle underlying each of these steps is that, ideally, each component of the Theory of Change is built on a foundation of strong community and stakeholder engagement and buy-in, as represented by the violet bar at the bottom of Figure 1. This not only includes

the data stewards and methods experts identified in other steps, but also people representing Hispanic/Latinx communities. CHI's engagement with community partners, such as the Denver Racial Equity Council, is described in subsequent sections.

Data Disaggregation: A Toolbox of Approaches

As identified in the Theory of Change model, CHI conducted a literature review to understand available methodological approaches for disaggregating data. To distinguish between approaches, CHI classified these methods into those that link datasets and those that expand datasets.

Data linkage is the combining or merging of data across two or more sources. The objective is to establish a link, whenever possible, between a person's record in one dataset and that same person's record in another. Using methods, such as deterministic or probabilistic record linkage, allows for the integration of two or more sources into a single dataset. A third method — statistical matching — uses correlations to identify similar individuals when direct linking is not possible. Using these methods, researchers and other stakeholders can answer additional research questions that might not be addressed with just a single dataset.

In contrast, **data expansion** does not rely on establishing links or matches between individual records in two datasets. Rather, this set of methods is characterized by filling in or making data more robust to allow for better reporting and analysis. For example, multiple imputation uses iterative processes to fill in missing observations based on information already present in the dataset. Another method, called race bridging, is often used by the National Center for Health Statistics (NCHS) to create uniformity in race/ethnicity categories when classifications change over time. Researchers could use multiple imputation or bridging methods to predict a person's ethnic identity when it is missing or not collected.

Table 1 displays the six promising linking or expansion methods that CHI identified for disaggregating data. Selection of methods will depend on a user's individual circumstances and objectives, as well as the considerations discussed in the next section. This is not intended to be an exhaustive list. Additional discussion and examples of each approach are included in Appendix A.

Figure 1. CHI's Theory of Change for Disaggregating Data on Hispanic/Latinx Ethnicity

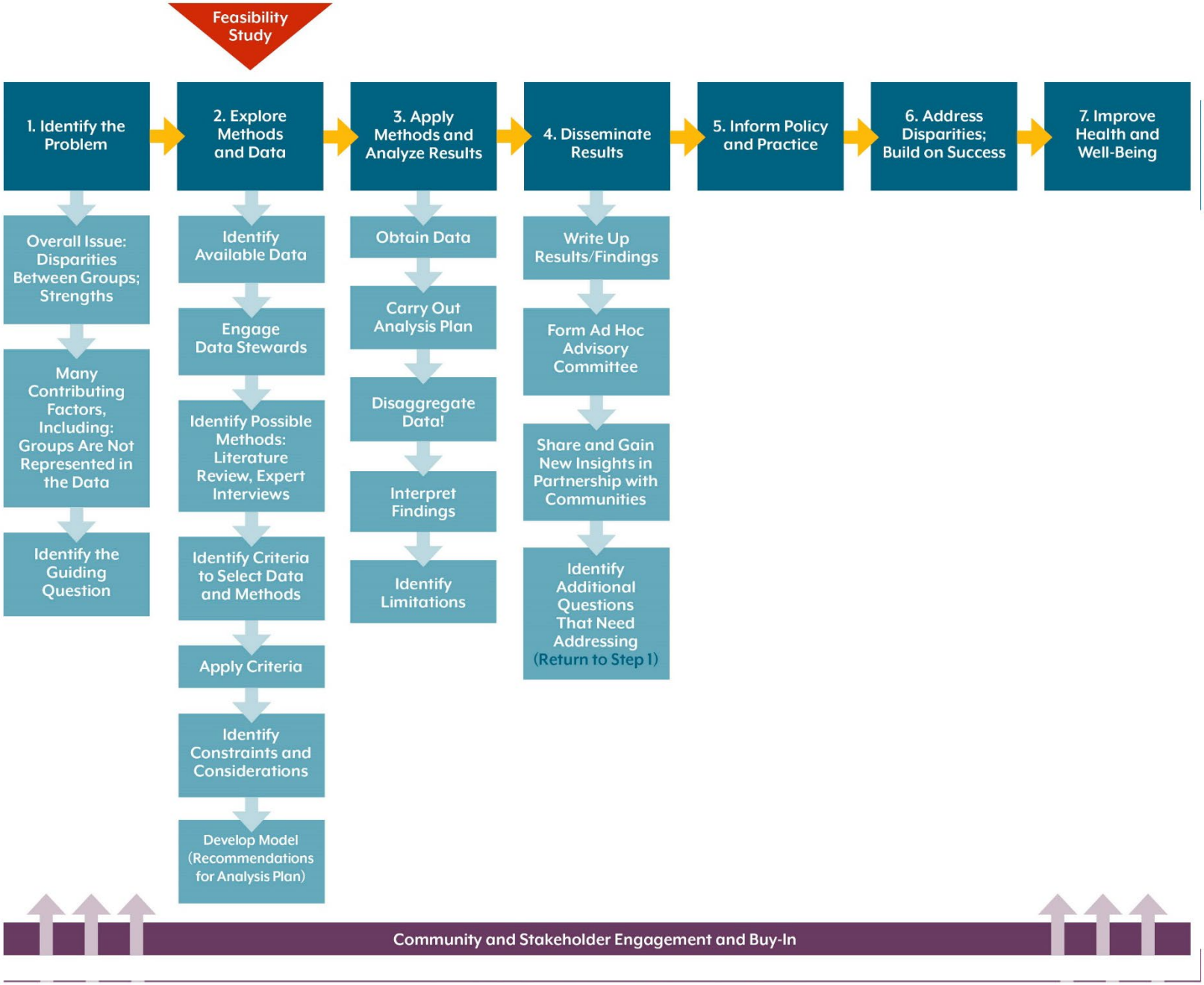


Table 1. Toolbox of Data Disaggregation Methods and Considerations[4,5,6]

	Methodology	Description	Methodological Considerations	What Questions Can the Method Answer?
Data Linkage	*Deterministic Linkage*	▪ Uses predetermined rules to pair records across datasets. ▪ Requires exact agreement on specific identifiers, such as social security number.	▪ Quick linkage method that is useful when records have unique identifier variables that are complete and accurate.	Linkage can expand knowledge on health indicators related to the same population of interest. By combining multiple datasets, additional observations of health outcomes and the impacts of identified factors on these measures could increase knowledge on specific topics of interest only found in certain surveys, health records, or other sources, like administrative data. *Example of Research Questions:* ▪ What is the impact of racial misclassification of American Indian/Alaska Native individuals in state cancer registries?[7]
	Probabilistic Linkage	▪ Matches duplicate records within or across files using non-unique identifiers such as name, date of birth, or address. ▪ Match weights are assigned to comparison pairs to estimate the likelihood that two records are a true match across multiple datasets, given the agreement across the identifiers.	▪ Error rates are usually unknown. ▪ Thresholds for the matching weights are subjective.	
	Statistical Matching	▪ Relies on correlations between variables shared across datasets by estimating relationships between variables (such as demographic characteristics). ▪ Estimates the likelihood that any given observations are similar, but not necessarily known to be the same individual between the two datasets.	▪ Conditional independence — which posits that the variables of interest are independent given the common variables in each dataset — is assumed between the variables of the two merged datasets. This assumption could introduce bias in analyses across the datasets if those variables are associated with each other. ▪ Auxiliary information can reduce uncertainty in the existing relationships between the variables of interest.	Statistical matching can expand knowledge of existing datasets, like surveys, that normally do not have personally identifiable information. *Example of Research Questions:* ▪ What is the relationship between household income and level of assets, after combining the Survey of Consumer Finances and the Current Population Survey?[8] ▪ What is the relationship between household income and patterns of consumption of goods and services?[9]

	Methodology	Description	Methodological Considerations	What Questions Can the Method Answer?
Data Expansion	*Multiple Imputation*	▪ Uses observed values from multiple iterations of a multivariate model to predict and replace missing data with a set of plausible values. ▪ Combines results from multiple iterations to account for variability.	▪ Relies heavily on assumptions because no records can verify that the outcome is correctly imputed; though tests can be developed using populated values to assess a model's predictive power. ▪ Assumes data are missing at random, allowing missing data to depend on observed values. ▪ Can introduce greater uncertainty if observed values are not strongly predictive of missing values.	Imputation addresses missing data issues by expanding existing datasets and creating more robust reportability for certain measures. *Examples of Research Questions:* ▪ How can the completeness of data be increased to make inferences about health outcomes across different race/ethnic groups from available claims datasets?[10] ▪ What are existing health care disparities in pediatric quality of care measures?[11]
	Bridging	▪ Creates population estimates for datasets where race/ethnicity classifications have changed over time or differ between data sources. ▪ Approaches the classification issue as a missing data problem.	▪ Auxiliary information is needed to understand population-based estimates for the percentage of individuals that fall into the desired race/ethnicity classification.	This method creates continuity across years of data available in certain data sources. To understand trends over time, it is important that individuals be classified comparably across different reporting schemes. *Example of Research Questions:* ▪ When racial classifications in the census change over time, how can we address issues in time trending?[12]
	Predictive Modelling	▪ Uses regression models to investigate the associations between variables of interest and an outcome of interest. ▪ The type of model needed depends on the outcome being investigated and will determine if the model is linear, logistical, multinomial, or a mixed methods approach, among others.	▪ Approach is based on the variables available to evaluate for significance to the outcome — as some datasets might be limited, this might also limit the predictive power of the regression model. ▪ Methods have been developed that use geographic and surname information for predicting race/ethnic classifications.[13,14]	In the context of data disaggregation, predictive modelling could assist in one of two ways: once disaggregated racial/ethnic groups data are available, associations between each group's data and their outcomes can be evaluated. A second use could assist in predicting the likelihood of an individual identifying as a member of a racial/ethnic subgroup. *Example of Research Questions:* ▪ How do rates of suicide deaths differ between racial/ethnic groups after mental health visits, and how accurate are these models?[15]

For more information and examples of the use of these methods, see Appendix A.

Key Considerations

To expand current data sources or linkage across multiple systems, several key considerations need to be addressed when investigating data disaggregation. The methodological development of linking across health data systems is complex, but dissemination of results and the impact they might have on stakeholders can also present its own host of issues. Considerations include:

Type of Data Available

To understand what methodologies researchers can use to expand knowledge and disaggregate data locally, a complete picture of the data must be considered. This includes gathering descriptive information of the data of interest to inform what variables can be used in the linkage strategies and if additional data management methods will be needed. Collaboration with the systems that house the data is important in this step to understand if it is feasible to access the data, as well as what additional resources may be needed to employ a particular strategy.

Other important considerations include cost of datasets, what years or months the data are available, administrative requirements to obtain data, the length of time for approval to obtain research files, and adequate documentation on the data of interest.

Privacy Concerns

Linking records with protected health information across multiple systems creates privacy concerns. Restrictions or laws may be in place that do not allow for the sharing of certain health or personal information across systems without informed consent of the patient. Such restrictions could limit the ability of users to link datasets using unique identifiers. Also, institutions might not have the infrastructure to securely house data with personally identifiable information.[16]

Even deidentified data could be of concern as it relates to analyzing and releasing results from disaggregated data approaches. Some unintended harm could occur when releasing information at a more granular level, as some areas could have unique individuals living in smaller populations. This requires balancing the desire to provide detailed information on communities with appropriately protecting information that might be identifiable because of its uniqueness.[17]

Limitations of Approaches

Each methodology will have limitations to consider when attempting to disaggregate data. Some of these limitations are identified in Table 1; further limitations are discussed in Appendix A. Being transparent about limitations in these approaches will be important to create trust among those who are sharing the information and those who are receiving disaggregated data.

Which Approach to Use?

Each organization pursuing data linkage and expansion techniques will have its own individual needs based on their available data systems. CHI developed a set of criteria to choose an appropriate approach when considering Colorado's data sources. Researchers, decision makers,

and other stakeholders can use this framework when considering an approach for their own analyses. These criteria and how CHI addressed them are displayed in Table 2.

Table 2. Addressing Criteria in the Context of Colorado's Data Systems

Criteria	CHI's Application of Criteria
What questions need to be answered with this approach?	CHI seeks to understand how Hispanic/Latinx subgroups differ on measures of access to care, use of services, and health conditions. To answer this question, we are assessing whether data disaggregation is possible in Colorado across multiple systems. Because of this, we plan to explore linkage and expansion strategies.
Do available systems contain data that identify unique individuals?	The CHAS is the source of our disaggregated ethnicity data. It does not contain identifying information, so we must consider approaches that do not require unique identifiers. This rules out deterministic linkage methods and introduces constraints on the use of probabilistic linkage approaches.
What is the outcome of interest?	Our outcome of interest is the disaggregated Hispanic/Latinx subpopulation ethnicities available on the 2021 CHAS.
Are there common variables between the different datasets?	The common variables that we can use in the model between datasets include race, insurance type, language, and age.
Do time frames of data match, and if not, are analyses still possible?	There is some overlap. The 2021 CHAS was fielded in spring 2021 but asks utilization questions about the 12 months prior to the survey. We could (eventually) obtain overlapping data for 2020 from CHORDS and CIVHC.
Does answering the research questions require merging data between systems, or should the current single data source be expanded?	Our research question could be addressed by both methods. To disaggregate health outcome data, however, we need to include a merging step or additional collection of these data. The absence of unique identifiers limits the capabilities of more classical linkage approaches.
Is institutional knowledge of methodologies available in house?	We will need to consult on additional statistical assistance on complex methodology approaches.
Are resources available to carry out the intended strategies?	CHI will need additional funding to purchase data from Colorado's data stewards, as well as additional support for the development of specific survey weights and statistical techniques to trend the CHAS over time due to new race/ethnicity options.

Based on these criteria, CHI created a proposed modelling approach that will explore expansion methodologies of **bridging** and **predictive modelling**, which are strategies that can be employed when unique indicators are not available.

After identifying these two methods, CHI pinpointed elements necessary to construct a proposed analysis plan for disaggregating data on Hispanic/Latinx ethnicity. These elements are described in the Colorado use case.

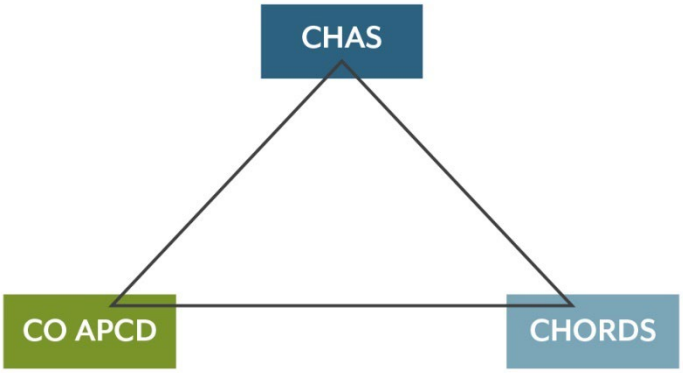
A Colorado Use Case: Using the CHAS to Expand Existing Data Availability

After reviewing literature and applying the identified criteria, CHI proposes an approach to expand Colorado data sources to disaggregate race/ethnicity information for the Hispanic/Latinx community. To do this, both bridging and predictive modelling techniques will be pursued. This Colorado use case explores the current data landscape in the state, outlines recommended phases of analysis, and identifies specific considerations and lessons learned to date.

The Current Data Landscape in Colorado

Three key health-related data sources are available in Colorado that could be used to disaggregate data for those who identify as Hispanic/Latinx. These three sources include the CO APCD, CHORDS, and the CHAS (see Figure 2).

Figure 2. The Colorado Data Environment and Linkage Opportunities



The CO APCD is the Colorado claims database housed within the Center for Improving Value in Health Care (CIVHC).[18] The CO APCD includes claims data on insured Coloradans from payers, including Medicare, Medicaid, and commercial insurance. The CHORDS system is a network of providers located primarily along Colorado's Front Range that brings together electronic health records to inform public health research and policies.[19] Because CHORDS is a network of health systems and providers, data in this system represent the care-seeking population.

The CHAS, a product of CHI, is administered every two years and captures data on health care access, utilization, barriers to care, and social factors in Colorado. In 2021, the CHAS expanded the number of race/ethnicity categories asked on the questionnaire to understand how these measures differ among specific communities in Colorado.

Additional information on each of these systems is included in Appendix B.

CHI's Proposed Modelling Approach

CHI considers race bridging and predictive modelling methodologies the most promising strategies for disaggregating data on Hispanic/Latinx ethnicity. The following phases of the analysis align with the components of Step 3 in the Theory of Change model.

Obtaining Data

CHI must build in time and resources to request research files from both CHORDS and CIVHC. Obtaining raw data files from the CHORDS network requires committee review of the proposal before a special custom query of the data from all the network's partners can occur. Based on talks with CHORDS experts, CHI expects that the CHORDS request will take about two months to fulfill. Similarly, CO APCD data covering calendar year 2020 will be available from CIVHC in July or August 2021. Delivery of the CO APCD data files is expected to take three months, based on discussions with partners. The CHORDS and CO APCD requests will need to be coordinated to create a long enough period between request and review to have data available before the end of the summer if the project is to be finished by December 2021.

The 2021 CHAS is in the field until late June or mid-July 2021. Once data are received from the survey vendor, management and cleaning will take place to create a research file that can be used for the purpose of applying expansion techniques. Cleaning and management will take place before data are obtained from the other systems. A preliminary investigation of the disaggregated data variables will occur to examine sample size and other analysis considerations, such as the application of an imputation step to account for missing data issues.

Additional activities that CHI will complete during this phase include:

- Identifying key statistical experts who can advise on development and analysis of the proposed modelling methodology. The CHI team has already made connections through the UCLA network, the CHORDS network, and the University of Colorado Anschutz Medical Campus.
- Exploring whether CHI will need additional statistical analysis software to complete predictive modelling and race bridging approaches beyond the capabilities of CHI's current software, SAS® 9.4.
- Exploring with the CHAS survey vendor the development of an alternate weight to account for the new disaggregated race/ethnic group data for use in future analysis of the CHAS.

Proposed Timeline of Obtaining Data: June–September 2021

Analysis Phase

Once data are obtained from CIVHC and CHORDS, analysis of these datasets will begin. This phase of work will require multiple rounds of contact between the statistical experts identified during the preparatory phase and CHI team members. Steps that need to occur before building the analysis models are:

- Conduct descriptive analysis of all three datasets to understand the percentage of missing data in each source.

- Compare data dictionaries available from CIVHC and CHORDS to the contents of each variable of interest.
- Harmonize the data sources, which includes data cleaning and management of variables of interest to transform them for use in the modelling approach.
- Investigate the collinearity and joint distributions of variables of interest to be included in the modelling approaches.
- Investigate if it is necessary to build imputation steps into the CHAS, CHORDS, and CIVHC datasets to expand available data on Hispanic/Latinx ethnicity and other race/ethnicity variables based on missing data analyses.

Once data files are prepped and ready to use, CHI will pursue the two identified approaches to disaggregate data. These approaches are:

- **Race bridging** to expand data currently available on the CHAS to allow for comparable reporting across years of data.
- **Predictive modelling** to investigate predictability of variables of interest and Hispanic/Latinx subgroups using the CHAS and applying the model to the other two data sources.

Race Bridging

CHI anticipates using the race variable in the model to disaggregate Hispanic/Latinx ethnicity. After CHAS data are received from CHI's survey vendor, CHI will create an analysis file. Part of this step will require a race bridging approach, as CHI added a new race category, Middle Eastern or North African, to the existing race question, an update from previous survey years.

To make estimates comparable over time, as well as to provide harmonization between racial groups available across all datasets in preparation for predictive modelling approach, CHI will need to bridge race categories from past surveys, reallocating counts into the previous classifications used on the CHAS. In other words, CHI will use this approach to predict how people identifying as Middle Eastern or North African would have answered the question if that category had not been asked.

Data Cleaning and Management

CHI will complete descriptive statistics of the available CHAS data to understand the level of missingness in the new disaggregated race categories. An imputation step will be used to address missing responses to assist in data analysis and reduce uncertainty due to these missing values. CHI will also explore how to handle respondents who identify with more than one disaggregated race category.

Development of the Bridging Model

Researchers have developed statistical packages for bridging in past analyses. CHI will also consider the use of this software to accomplish bridging. One specific package, MICE R 3.1.1, may be a statistical software approach to facilitate race bridging.[20,21] Another approach could include the use of IVEware software, available from the University of Michigan Institute for Social Research. This software includes SAS macros that have a similar function to the MICE package available in R.[22]

For the bridging technique, CHI will build regression models to relate individual-level and county-level covariates. CHI will also investigate the following individual-level characteristics as variables in the model from the CHAS: sex, Hispanic/Latinx ethnicity, geography, race, education level, age in years, insurance type, income, and language.

At the county level, CHI will include auxiliary information from the U.S. Census Bureau's American Community Survey as contextual covariates. Contextual covariates of interest include urbanicity, reported ancestral or ethnic origin, and the percentage of each county's population that reports more than one race. County-specific, single-race population percentages will also be included. Regression coefficients generated from the bridging models will be used to generate probabilities of selecting each possible primary race for respondents from the race group categories on the 2021 CHAS. This approach will follow methods as outlined by Thompson et al. and Schenker et al., as well as methods used by the NCHS to address classification schemes in the 2000 census.[23,24,25]

Validation of the Bridging Model

CHI will evaluate the quality of the bridging model the reliability of the methods. To do this, CHI can use an approach that tests the model on known ranges of data. By randomly deleting known values and running the model on these ranges, CHI will be able to understand how well the model predicted the ethnicities under Hispanic/Latinx identity. Model diagnostics will also be employed to understand the predictive performance of the race bridging approach.

More information on the race bridging approach can be found in Appendix A.

Predictive Modelling

After bridging is complete, CHI will use predictive modeling to estimate the likelihood that an individual is a specific sub-ethnicity within the Hispanic/Latinx population on the CHAS. For discussion purposes, this example uses the identity of Chicano.

CHI will build a regression model based on the CHAS to predict the likelihood that a person identifying as Hispanic or Latino* also identifies as Chicano. CHI will then apply this model in both the CHORDS and CO APCD using common variables across the datasets. An in-depth analysis of common variables and predictive power of those variables will be required, as well as analysis of the disaggregated data from the CHAS to discern the sample size and feasibility of exploring this technique.

Harmonization of Data Sources

In addition to the race bridging approach described above, there are several additional steps that need to take place to ensure data are comparable across the three data sources. Harmonization steps that CHI has identified include:

- Investigating patterns of response selection among individuals categorized as multiracial.
- Imputing values for the current Hispanic or Latino ethnicity variable available across all datasets to address missing data and to get a robust denominator population.

* The CHAS survey instrument uses the language Hispanic or Latino when referring to this group as an ethnicity. When referring to the variable on the survey, "Hispanic or Latino" is used.

- Exploring all common variables to be included in the regression model (see Table 3) and combining categories within these variables as needed to make them comparable.
- Exploring reporting patterns for those who indicate Hispanic or Latino as their race on the CHAS survey and possibly applying a race bridging approach to categorize them into one of the standard racial groups (refer to Appendix B for additional information on how the Hispanic or Latino ethnicity question is asked on the survey).

Analysis of Common Variables

For this step, CHI will investigate the statistical relationships between the available common variables and the outcome variable, outlined in Table 3. This step will also evaluate the number of respondents who identified disaggregated racial/ethnic subgroups and the distribution among the options available.

Table 3. Common Variables Available for Model in Colorado Data Systems

Common Variables Across All Systems	Additional Common Variables Between CHAS and CO APCD	Additional Common Variables Between CHAS and CHORDS
• Race • Age • Gender • County of residence • Insurance coverage type	• Disaggregated ethnicity • ZIP code	• Language

Development of the Regression Models

The outcome of interest and common variables will be included in the development of the models. After the analysis of the common variables and their predictive power is identified, those selected variables will be included in the models. Because some common variables will differ between CHAS and the two other datasets, additional models with the expanded common variables will be investigated to understand predictive power of these models — one for the CO APCD and another for CHORDS. The approach, though, will seek to identify a generalizable model.

The outcome of interest will be the disaggregated categories of the Hispanic or Latino ethnicity variable. Because of the complex nature of this approach, CHI will begin by creating a dichotomous outcome of one of the disaggregated categories of interest to understand ability to predict race/ethnic outcomes on the CHAS. The distribution of the disaggregated categories will be important for this step — if all Hispanic/Latinx respondents answer one specific identity, this may call into question the ability to disaggregate further within this population. Much of this approach depend on sample size of the disaggregated groups and how well they are represented in the available data.

Common variables to be included as covariates will include age, race, sex, geography, insurance type, as well as the additional common variables identified in Table 3 for each dataset of interest for the additional models evaluated. Because the outcome of interest is a dichotomous outcome, a logistical regression approach will be used. Model building will apply SAS® 9.4 as the statistical software package.

Quality Assessment

A quality assessment step will be built in to understand the predictive power of the regression model. Model diagnostics will be evaluated using performance statistics to understand the predictive performance of our model. These could include statistics such as the area under the receiver operating characteristic (ROC) curve to understand how much of the variance in the outcome is explained by the predictors in the model.[26] Some disaggregated data are also available from some providers in the CHORDS network and some payers in the CO APCD. These data will be used as a comparison population to understand how well the model predicted the likelihood that someone identified as the disaggregated race/ethnic subgroup. This approach uses two datasets:

- The *training dataset* is the dataset on which the model (to predict disaggregated Hispanic/Latinx ethnicity) is built.
- The *validation dataset* contains disaggregated data. For example, this could be the subset of CHORDS or CO APCD data that already contains disaggregated Hispanic/Latinx ethnicity data.

By comparing the actual values to the predicted likelihood values in the model, CHI will be able to determine how well the regression model predicted the ethnicities under Hispanic/Latinx identity. Another important element to validate the model will be the identification of a survey or additional auxiliary information that has the available distribution of the disaggregated race/ethnic subgroups in Colorado. This information will be important to understanding whether the CHAS, CO APCD, and CHORDS are representative of these subpopulations.

More information about predictive modelling can be found in Appendix A.

Through both race bridging and predictive modelling methods, CHI will explore the analytical approach of disaggregating data and identify a dissemination plan to report results from these analyses. By combining these two approaches, CHI will be able to expand existing analytic ability within the CHAS as well as understand a linking strategy between Colorado's data partners.

Proposed Timeline for Analysis: September 2021–February 2022

Stakeholder Engagement and Sharing of Results

As described in this project's Theory of Change model (see Figure 1), engagement of community stakeholders who are represented in the disaggregated data is an integral step of advancing equity. Including in the research process those who stand to be affected by the analysis builds trust and attempts to counteract the legacy of research not being sensitive to the needs of communities at best and exploiting or harming communities at the worst. Effective engagement also avoids the researcher "parachuting" into a community, collecting the data, and leaving the community without offering anything in return. In addition, engagement should not only focus on deficits, but also highlight the strengths within communities.

In the next phase of work, CHI proposes forming an ad hoc advisory committee of people who represent and serve Hispanic/Latinx Coloradans, including researchers, providers, advocates, and community leaders. This group will inform future phases, including sharing the results.

The dissemination step of sharing initial findings and results is particularly important in:

- Interpreting the data
- Understanding contextual or confounding factors
- Identifying limitations, considerations, and unintended consequences
- Sharing in data ownership
- Identifying additional questions and potential analyses

The underlying principle guiding dissemination is that it should be a two-way street: researchers share insights from the data with the communities involved, and the communities share their insights with attentive researchers. This relationship requires trust, openness, and a commitment to identifying ways in which the communities, first and foremost, benefit from the analysis.

CHI anticipates that there will still be some groups of Coloradans for whom the analysis will not yield reliable results. Community engagement allows for collection of additional data — such as stories, other qualitative information, and additional quantitative data that individual communities have collected.

CHI's proposed dissemination strategy will start with individual outreach to leaders of Colorado-based Hispanic/Latinx organizations. These may include Servicios de la Raza, Colorado Latino Leadership, Advocacy, and Research Organization (CLLARO), the Latino Community Foundation of Colorado, Lake County Build a Generation, Tepeyac Community Health Center, and Mountain Family Health Centers. This individual outreach will offer the opportunity to explain the potential benefits of the data disaggregation analysis and understand applications and drawbacks.

Additional dissemination strategies include publishing blogs and briefs in both English and Spanish; making presentations to policymakers and key stakeholder groups; and promoting the findings on social media. One group that CHI has identified as a future partner in this phase of the work is Denver's Racial Equity Council spearheaded by the Mayor's Office of Social Equity and Innovation. This group has worked with CHORDS partners in the past and has expressed interest in data disaggregation as a mode of advancing equity in Colorado.

Proposed Timeline of Dissemination: February 2022–April 2022

Limitations and Additional Considerations

There are some limitations and considerations that CHI will need to address. Because of the addition of the disaggregated race/ethnicity identities, CHI will need to consider weighting strategies in the CHAS so that estimates are not biased due to sampling.

In addition, the changes in classification schemes will create additional barriers to analysis, as CHI established a new categorization system by adding Middle Eastern or North African to the main race/ethnicity survey question. Understanding how this change will affect the ability to report estimates over time will be an important component when applying the race bridging technique. Considerations for this include:

- How might the individuals who identified as Middle Eastern or North African have reported in the past?

- How can we use auxiliary information to address changes in classification schemes?

The proposed predictive modelling method creates additional consideration for the next phases of work. These include:

- Do the available common variables have enough predictive power in the models?
- Is there a dataset available that has the disaggregated subgroups of those who identify as Hispanic/Latinx to validate the model findings?
- Is there a large enough sample size of individuals reporting the disaggregated race/ethnic subgroups to get a representative sample for analysis?
- How might CHI support data collection methods in CIVHC or CHORDS to address missing data issues?

Disaggregating data presents a unique consideration as it relates to data confidentiality and disclosure risk. To mitigate any issues in this area, CHI will consider the follow items:

- Which stakeholders are important to engage to understand important issues around data disclosure for specific groups in Colorado's communities?
- What are steps CHI can take in the dissemination process to avoid issues with disclosure of potential identifiable information for some subgroups within Colorado's population?
- What steps do CHORDS and CIVHC take to reduce risk of identifying individuals in their own data systems? What lessons can be learned from these data experts?

Other limitations may be the time frames available across all data sources. As the CHAS was fielded in early 2021, there may be some overlap between the CHORDS and CO APCD systems, but there is a lag in reporting for both claims and electronic health record data. CHI must coordinate with the data systems to understand how long it takes to get comparable points in time to the CHAS. This may present issues in the proposed timeline for when CHI can obtain the data.

Lessons Learned

The feasibility assessment provided important findings surrounding engagement and outreach to community group and data experts. To be successful in disaggregating data, collaboration is a required step to bring all stakeholders to the table to understand the feasibility of existing approaches. Lessons that CHI learned from this process follows.

Engagement with Data Stewards and Other Experts

Engagement with experts is integral in developing sound partnerships for the investment in future work in service of data disaggregation. Because of this engagement, CHI learned that:

- It is important to identify a champion within each data steward organization to create partnerships to further future work. Champions at both CIVHC and CHORDS provided detailed information about the data systems available, as well as the processes necessary to request data in a timely manner for additional phases of the data disaggregation work.

- Investment by the data stewards and identified champions is integral in developing plans for future phases of work. These champions are interested in the approach CHI is pursuing and are on board for future phases of work to pursue the project's goals.

CHI also engaged with statistical experts to understand the complexity of data linkage and expansion methodologies. The UCLA Center for Health Policy Research coordinated the introduction of one statistical expert with an existing relationship at the university. CHI identified additional support locally at the University of Colorado Anschutz Medical Campus. Through these engagements, CHI learned that:

- Statistical methodologies and their applications are very complex and may require additional expertise to apply them. Based on our institutional knowledge, it will be important to build in additional resources and time to collaborate with experts when the proposed methodology is pursued.

Because these experts were identified early on, CHI has a better understanding of the systems and methods in which data can be linked and expanded in service to data disaggregation. Early engagement also allowed for collaboration between CHI and champions in each data system. This provided opportunities for these champions to become invested and supportive of the proposed plan moving forward.

Engagement with Community Groups and Other Data Users

CHI plans to engage with community stakeholders throughout the phases of analysis and dissemination of results from the analysis. Outreach to stakeholders began in the planning for the 2021 CHAS questionnaire. As part of this feasibility study, CHI added a new item inquiring about specific ethnic identities among respondents who identify as being of Hispanic or Latino ethnicity. In addition, CHI added a Middle Eastern or North African response option to the race question and new ethnicity subgroup questions for all race options except white. A tribal affiliation question for those identifying as American Indian or Alaska Native (AI/AN) was also explored, but disaggregated groups were ultimately not added after discussions with stakeholders. See Appendix B for further discussion of the additions to the CHAS.

In developing the questionnaire, CHI reached out to organizations in other states for models of how the questions could be asked. The team also connected with Colorado-based organizations that serve communities of color. These included:

- Asian Pacific Development Center
- Center for African American Health
- Centers for American Indian and Alaska Native Health, Colorado School of Public Health
- Denver Indian Health and Family Services
- Lake County Build A Generation
- Latino Community Foundation of Colorado
- Mountain Family Health Centers

In general, stakeholders representing Middle Eastern or North African, Black or African American, Asian, Native Hawaiian or Pacific Islander, and Hispanic/Latinx communities supported adding the new ethnic identity items specific to their group. They appreciated the potential of additional insights gained about subpopulations.

The most significant pushback came from proposing a question on AI/AN tribal enrollment or affiliation. The lessons learned from this experience include:

- Data sovereignty — including how data are collected, stored, shared, and distributed — is an emerging topic for many AI/AN tribes/nations. Sufficient time and resources are required to consult the tribes/nations about how the data will be used.
- How different communities, and in this case, people who identify as AI/AN, are reflected in data reporting must be carefully considered. Researchers may encounter backlash if a particular tribe was not consulted prior to data collection or reporting if it reflects poorly on the tribe.
- Tribal enrollment and tribal affiliation are two different concepts. On a survey, it can be difficult for people to answer if they affiliate with multiple tribes but officially are enrolled in one of them.
- Tribal enrollment or affiliation can be considered personally identifiable in small sample sizes. Colorado only has two recognized tribes.

The primary lesson learned was that trust needs to be built among many communities given historic racism, exploitation, and violence. This takes an investment of time.

Conclusion

Disaggregating data in Colorado's current data systems presents an important opportunity to identify existing health disparities and hidden strengths among local communities. Colorado's claims data housed within CIVHC and electronic health record data available through the CHORDS network are two available systems that could be expanded from such an approach. Disaggregated race/ethnicity data from the CHAS could assist in the process of creating such results through methodological approaches of race bridging and predictive modelling.

The entire process of engaging these systems in service to disaggregating data provides a distinct opportunity for collaboration among Colorado's stakeholders. Partnerships between CHI, CHORDS, CIVHC, and other community stakeholders will foster future interest in identifying health outcomes among specific communities of color in the state.

Appendix A: Profiles of Methods Featured in the Toolbox

Deterministic Linkage

Deterministic record linkage is a method used to link across databases when unique identifiers for individuals are available. An example of this is a unique medical identifier used across the same health system to identify patients. Researchers may also attempt to match unique individuals in multiple datasets by using more than one identifier, including social security numbers, patient record numbers, birth dates, or first and last names of the person if these elements are available across all datasets. Using multiple identifiers can help match across multiple datasets more precisely when certain variables, like social security number, are not available.[27]

Typically, deterministic linkage requires that there be exact agreement across the unique identifiers used. Misspelled names and other errors can cause a high degree of missed links that keep the variables from agreeing with each other. Some methods, like a stepwise method, allow links to be made if all but a certain number of variables are linked properly across the datasets.[28]

Deterministic linkage methods are quick and very useful when records are complete and known to be accurate.[29] However, because many sources in public health surveillance and administrative data are not always complete and have a degree of data entry error, deterministic linkage is not the most appropriate method in these instances.

One of the best-known examples of deterministic linkage came in 2012 when the National Center for Health Statistics (NCHS), the Centers for Medicare and Medicaid Services, the Social Security Administration (SSA), and several other federal agencies came together to link the NCHS data systems and Medicare files.[30]

Probabilistic Linkage

Because of these issues associated with deterministic linkage, probabilistic linkage approaches have been developed to link data sources that do not share unique identifiers. Probabilistic record linking instead calculates a probability that two records in different data sources are associated with the same individual.[31] Probabilistic linkage uses identifiers such as first name, last name, date of birth, or address. When these are all used together, these identifiers may uniquely identify a person across data systems.[32]

Probabilistic data linkage has several key steps in the pre-linkage, linkage, and post-linkage stages. Before methods are used to link data, data must be standardized. The success in the linkage of data depends heavily on the quality of the data used. The first step is thorough data cleaning and management. Matching variables then need to be selected, which depends on the contents of the datasets of interest. When choosing matching variables, the uniqueness, availability, accuracy, and stability over time should be considered. For example, personal characteristics, like ethnicity, date of birth, and place of birth, are fixed and rarely change over time, while social demographic variables, like marital status and address, are subject to change over time.

Other steps, like the selection of blocking variables, such as geographic region or specific clinical condition, should also be considered during this phase. Incorporation of blocking variables avoids comparison of record pairs that are the least likely to be matches across datasets, which can reduce the amount of time it takes to match one-to-one pairs.[33]

During the post-linkage stage, it is important to understand the efficacy of the linkage. This process includes evaluating the sensitivity of a link (correctly matching true pairs), specificity (correctly not linking non-matches), and precision (correctly matching pairs out of all possible links). These measures help researchers understand the quality of the linkage and are always important to build into the linkage approach.[34]

There are many applications for probabilistic record linkage. Early applications of this method were used to update and maintain large national health and death indexes and compare these metrics across multiple sources. Probabilistic linkage can also be used to reduce the number of duplicated records in a national registry or within a survey frame, potentially reducing bias in sampling.[35]

In more recent years, probabilistic linking methods have been used to expand data collection across multiple sources to answer additional research questions that are not possible with only one of the sources in question. An example of probabilistic linkage can be seen through methods developed to link data between the National Trauma Data Bank® and the Traumatic Brain Injury Model Systems to expand opportunities to explore additional data for individuals affected by traumatic brain injury.[36,37]

Statistical Matching

In many cases, data that are available to researchers are in the form of public use files that have had unique identifiers scrubbed from the datasets. Statistical matching, also known as data fusion, data merging, or synthetic matching, is a model-based approach to combining existing surveys from the sample population when unique identifiers are not available. The application can be beneficial in many ways, as it does not require fielding an additional survey to get more information on the same population and will not require extra costs as a result. One way it can be applied is to create a new data file where all data are available as variables across the different datasets, and where records are generated using information across common variables between data sources.[38]

Unlike record linkage, where the method deals with identical individuals, statistical matching deals with similar individuals. One data source is used as the recipient file in which new information is imputed for each record using data from a separate source as the donor file. Statistical matching differs slightly from imputation, however, in that there are additional variables of interest that are not available in the recipient file. In this method, data are brought together from both datasets and are based on different units of measure, creating a new dataset that expands across both linked files.[39]

Statistical matching has a distinct limitation that surrounds the conditional independence assumption, an important concept for probability distributions over multiple variables. For example, if one dataset contains two variables (A, C), and another dataset contains the two variables (B, C), the conditional independence assumption tells us that the two variables, A and

B, are independent given the third variable, C. However, this assumption would not be testable if variables A and B are not observed jointly on the same individuals.[40]

This condition is important to statistical matching because the purpose of matching across multiple datasets is to analyze the joint relationships between these variables A, B, and C. If the true relationship between these three variables is not true to conditional independence, then this assumption would mask the real relationships, thus introducing bias into the analyses of these variables.[41]

Statistical matching procedures are complex and require several key steps to successfully merge data sources. The first and most important step is that of data harmonization. To merge two surveys, the sources must match based on common definitions of the variables of interest and reference period of interest. Harmonization also identifies the need to adjust for any missing data in key indicators as well as adjusting for any measurement errors in the datasets. Once this step occurs, all common variables need be analyzed to understand their distributions. After descriptive analysis of common variables, a statistical matching method must be chosen. There are many to choose from, including hot deck methods, regression-based methods, mixed methods, or multiple imputation methods, among others. Understanding the source datasets and any constraints these impose will determine what approach to pursue.[42]

Examples of statistical matching methods that have been used to create synthetic datasets for health research include bringing several measures of quality of life into a single data source and bringing together information on employment and income.[43]

Metrics on the quality and precision of the merge are important when evaluating the statistical matching procedure. It is important to evaluate uncertainty to assess the conditional independence assumption as well as understand the distributions of variables from the original datasets and the matched dataset.[44]

Multiple Imputation

Imputation is a commonly used technique that addresses missing data in the observed variables of datasets. The method creates several different plausible datasets and combines results obtained from these simulations.[45] Multiple imputation is used to replace missing values with a set of values generated from observed distributions. For each missing value, the distribution of the value is computed based on other variables in the model. The method assumes every missing value is missing at random.[46]

In the first step of imputation, multiple copies of the dataset are created, with missing values replaced by imputed values. These values are obtained from the predictive distribution of data in the dataset. The imputation procedure accounts for uncertainty in the estimates by inserting variability into the imputed values since the true value of the missing data is unknown. After this iterative process, a statistical model is fit to each of the imputed datasets. Associations among variables will differ among the datasets because of the variability introduced by the imputation procedure — these are then averaged together to give an overall estimated association in the model.[47]

While imputation methods can expand known datasets, several considerations should be addressed. Instances can occur where variables included in the model are not normally

distributed, which may introduce bias. Another issue is the assumption that data are missing at random, which might not always be the case. Unfortunately, the magnitude of the problem that the administrative collection of these data has on the impact of a model is difficult to know. Possible reasons why these data are missing must be considered, along with the likelihood that missingness is dependent on some factor. One method could be comparing the individuals with missing data to those with available data and evaluating any patterns that might emerge between these two groups.[48]

A multitude of examples of imputation methods exist, since it is commonly employed to adjust for missingness. In one example, multiple imputation was used to address missing values after linkage between the NCHS data and Medicare enrollment and claims records.[49] In another example, multiple imputation methods were employed to impute race/ethnicity data using Medicaid enrollment data.[50] Researchers evaluated methods used to impute race/ethnicity data, providing analyses and procedural examples of different ways to approach multiple imputation methods to address the missing data problem.[51]

Bridging

Race bridging was developed as a method to recategorize individuals when race/ethnicity classifications are updated, which has happened several times on the U.S. Census. As updates make it difficult to compare estimates for different race/ethnic groups over time, bridging can be employed to model and predict race/ethnic groupings as these classifications change to help with continuity of reporting.[52,53]

These methods were developed in response to changes in multiple race reporting. In 1997, the Office of Management and Budget (OMB) revised its directive on the collection of race/ethnicity data in federal systems, allowing respondents to choose more than one race. It quickly became apparent that approaches to make data comparable over time were necessary, giving rise to the bridging methodology.[54] Essentially, race bridging allowed OMB staff to predict how people would have identified their race if selecting only one race.

An example of this method was employed by Parker et al. With data from the National Health Interview Survey — which has allowed respondents to select more than one race group prior to the 1997 update on federal standards — researchers used the bridging approach to model and predict primary-race categories for each multiple-race that was reported by respondents. In this example, separate logistic regression models were developed for each two-race group with a large enough sample.[55] This example uses newer classifications to categorize respondents into a previous reporting system.

In a more recent example, Thompson et al. applied a forward bridging approach to address classification changes in the Asian race/ethnicity category on death certificates. Between 2003 and 2011, states began to adopt an expanded number of categories under the Asian and Pacific Islander group, all on a state-by-state basis. Researchers approached these staggered adoptions as a missing data problem as well, employing the bridging technique to estimate the distribution of respondents as "other Asian or Pacific Islander" into the expanded classification system for data gathered before the adoption, which now includes Korean, Asian Indian, and Vietnamese subgroups.[56]

Much like other imputation methods, the bridging method approaches these issues as a missing data problem, where records are viewed as missing the alternative categorization method.

Predictive Modelling

Predictive models can include a variety of approaches, and these methods are constantly being updated and created anew. In this report, predictive modelling refers to the creation of regression models that analyze the association between predictor and outcome variables. Some common types of these predictive models include:

- Linear regression
- Logistic regression
- Poisson regression
- Polynomial regression

The type of modelling approach needed depends on the types of data that researchers and other stakeholders have available. The outcome or dependent variable of the analysis is the variable that researchers want to predict, using a host of independent variables, also called predictors. If the outcome of interest is dichotomous, nominal, or continuous, approach types will differ based on these data.[57,58]

Descriptive analysis of variables included in models is necessary to understand their distribution and to discover any additional pitfalls they might inject into the model. An example of this is the concept of multicollinearity. Multicollinearity occurs when two or more of the independent variables are highly related or correlated. Adding highly correlated predictor variables into the same model can decrease the precision of estimated regression coefficients and reduce the effectiveness of the predictive power of the model.[59] Variance inflation factors can be used to detect multicollinearity among the predictor variables.[60] Other considerations exist based on the type of model that is pursued. Paying attention to these considerations and developing strategies to adjust or address these concepts must be a part of the overall regression approach.

Predictive modelling and regression analysis is applied as a method in many different industries. These types of analyses are the basis of epidemiological and pharmaceutical research as scientists seek to understand the impact of certain factors on their research outcomes. Racial/ethnic data are frequently studied and developed into modelling approaches to adjust for and understand differences in health, access, diagnosis, and other outcomes where it is important to understand inequities that exist between groups.[61,62] One example illustrated how investigators analyzed differences in cancer outcomes across demographic characteristics, including race/ethnicity among patients with pancreatic cancer.[63]

Using regression models to predict race/ethnicity as an outcome is a more difficult task. Many approaches use both geographic and surname information within their sample to predict race/ethnicity, using approaches like the Bayesian Surname and Geocoding method or any of its adaptions.[64,65]

When surname data are missing, other demographic characteristics and available geographic data must be used to make inferences. Investigating these predictor variables and

understanding how different social, demographic, or environmental factors are correlated with race/ethnicity can be a tool to further understand how these variables interplay with race/ethnicity data in local communities.[66] One such example is the Pew Research Center's approach to estimating the number of immigrants without documentation living in the United States. Combining information from the American Community Survey and the Current Population Survey, researchers developed a model that predicted the likelihood that an individual was an immigrant without documentation, based on key factors associated with this population.[67]

Appendix B: Colorado's Current Data Sources

Center for Improving Value in Health Care and the Colorado All-Payer Claims Database

Description

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 Colorado All Payer Claims Database (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.

The process of how the CO APCD works is described as the claims lifecycle. Health care payers provide a data set that contains information on every covered plan member who is a Colorado resident whether or not the member utilized services during the reporting period. The file must include member identifiers, subscriber name and identifier, member relationship to subscriber, address, age, race, ethnicity, and other required fields to allow retrieval of related information from pharmacy and medical claims data sets. First, a health care provider submits a claim for payment to the health insurance company or other payers, like Medicaid or Medicare. The claim contains information on items like charges, diagnosis, location, and the services rendered during that visit. After the payment is processed, the insurance company submits the claims information to the CO APCD. CIVHC then processes these claims and turns this information into public and custom datasets about how Colorado is receiving care.[68]

State of Race/Ethnicity Data

The CO APCD collects data on both race and ethnicity across several fields. Robustness of current data reporting for race and ethnicity differs across the payers in the CO APCD. Among commercial payers, 81.7% of the race variables had missing or unspecified race information for patients in 2019, compared to 83.5% in 2020. For Medicaid in 2019, 8.2% were classified as "Not Provided," with another 1.4% classified as "Other/Unknown," but percentages improved in 2020, with 6.1% classified as "Not Provided" and 1.1% as "Other/Unknown." Medicare had more missing data, with 61.8% classified as "Unknown or Not Specified." Data for Medicare claims was only available for 2019.

Data available on Hispanic ethnicity had similar percentages of missing data, where 78.5% of commercial claims had "Unknown" Hispanic ethnicity, while Medicaid did not have any claims with a reported "Unknown" Hispanic ethnicity. Medicare fee-for-service reported 56.3% of claims with "Unknown" Hispanic ethnicity, while Medicare Advantage plans reported 49.9% of claims with missing Hispanic ethnicity data.[69]

The CO APCD has additional fields collecting race/ethnicity data for more disaggregated groups, but reporting is limited and much of the data are missing.

Limitations

One limitation of the CO APCD data system is that, because it is a claims database, only insured populations are represented. However, race/ethnicity data are based on member eligibility files, which means that even though insured populations are represented, these individuals are counted even if they did not access care.

Submission of Employee Retirement Income Security Act-based self-insured employer claims is voluntary, so most data from self-insured entities are not included in the database.[70] This could impact the generalizability of some of the estimates and may require limiting of the datasets when linking.

Prospective Developments

The CO APCD system currently has some disaggregated data on race/ethnicity available in its system. However, not all insurers are collecting this information and missingness of data is an obstacle. The Colorado Health Institute (CHI) plans to coordinate with CIVHC to understand the feasibility of integrating the disaggregated race/ethnicity information into collection efforts.

Maria de Jesus Diaz-Perez, PhD, Center for Improving Value in Health Care, contributed to this section.

The Colorado Health Observation Regional Data Service (CHORDS)

Description

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.[71,72]

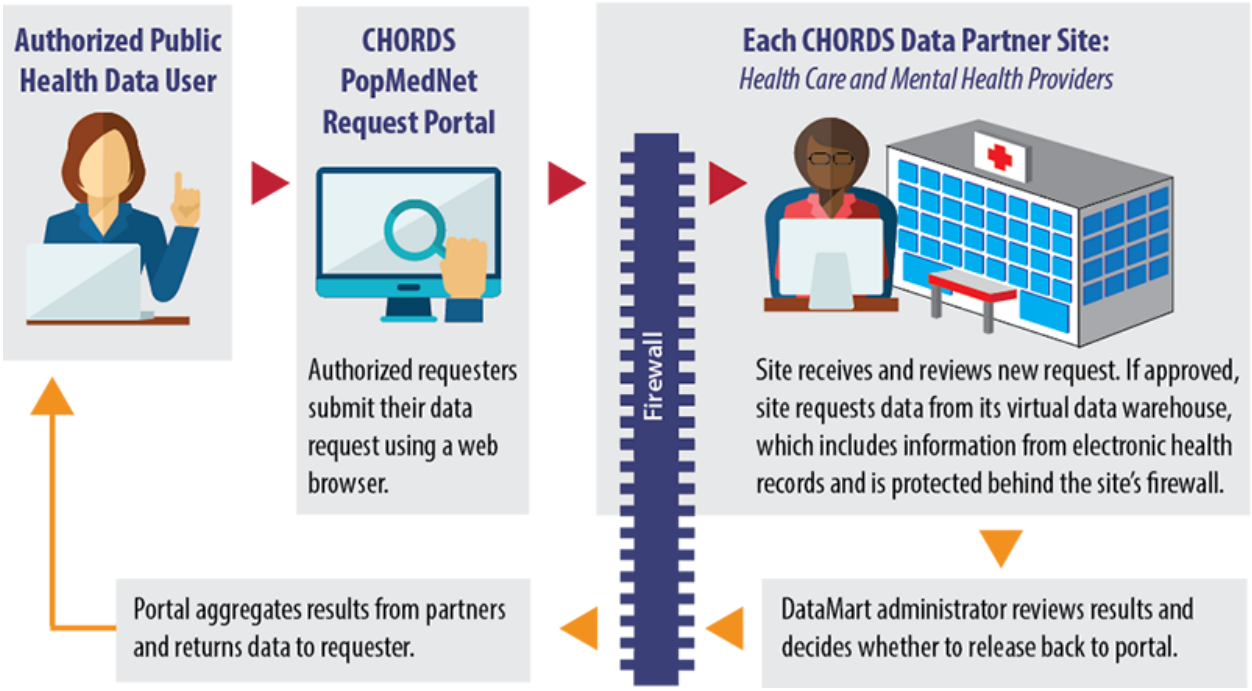
CHORDS uses a federated data model, which means that each data partner maintains autonomy over its data. Instead of loading data into a central repository, data are queried from each data partner separately and then aggregated within a secure environment (called PopMedNet). This enables health care data to remain in health care organizations' secure environments until data are required for a defined public health or research question.

All data partners conform their EHR data to standard tables and fields in a virtual data warehouse (VDW). Common components available from CHORDS include encounters, demographics, diagnoses, and prescriptions. Each data partner populates as many fields as possible within each table. CHORDS conducts routine quality assurance on each data partner's

VDW. The quality assurance reports assess each CHORDS table for data model conformance, data plausibility, data completeness, and data persistence.

CHORDS currently has about one million unique patients with at least one encounter for any given year. Each CHORDS data request creates a unique, limited dataset. Figure 3 shows the process of requesting CHORDS data in a federated model. CHORDS has developed a standard set of queries to produce estimates for common diagnosis-based health conditions, such as prevalence of diabetes, tobacco use, depression, or COVID-19 diagnoses. If a CHORDS data user is interested in data that are not part of a standard request, then they work with the team of developers to write a unique query.

Figure 3. The CHORDS Federated Model for Data Collection[73]



State of Race/Ethnicity Data

The CHORDS partners collect patient race/ethnicity data. There are five fields in the VDW that collect race data and one field that collects ethnicity data. Multiple fields give data partners the opportunity to include multiple races for a given patient. Sometimes a patient has an unknown race for the primary field, but a known race provided in the other collected fields. In these cases, the other available fields can be used to fill in missing information. All 14 of the CHORDS data partners populate the primary race field; five currently populate data in one or more of the additional race fields.

There are seven race categories for these fields, including: American Indian/Alaska Native, Asian, Black or African American, More Than One Race, Native Hawaiian or Other Pacific Islander, Unknown or Not Reported, and White.

CHORDS currently only captures Hispanic ethnicity. This field contains three possible categories: Hispanic, Not Hispanic, and Unknown Hispanic.

Based on a data request of the CHORDS network from the 2019 calendar year, 19.0% of individuals had a race designation as Unknown or Not Reported for partners representing the metro Denver region's seven counties. As for Hispanic ethnicity data, 9.4% of this care-seeking population were classified as unknown.[74] Understanding the breadth of missing data for race/ethnicity reporting will be important for CHI in future phases of work and coordination with CHORDS to identify ways to expand the current state of the data. Some approaches are identified in the Prospective Developments section.

Limitations

CHORDS data are mostly representative of the metro region of the state, which includes Adams, Arapahoe, Boulder, Broomfield, Denver, Douglas, and Jefferson counties. Data are also only representative of individuals seeking health care services and are not a random sample of the underlying population.

Prospective Developments

CHORDS has expressed interest in integrating disaggregated race/ethnicity data into the standard query. Some partners are currently collecting this information, but these data are not required for the standard reporting that CHORDS performs. Additionally, CHORDS data experts and partners are investigating the potential use of language as a method for identifying more disaggregated data on patients and the use of multiple imputation to expand the CHORDS dataset. These future developments are described in more detail below.

At least one partner collects some national origin data for Hispanic patients. National origin categories include: Cuban; Mexican, Mexican American, or Chicano/a; Puerto Rican; Other Hispanic, Latino/a, or Spanish Origin. These data are not currently available in the VDW. Whether other partners collect more detailed race, ethnicity, or national origin data is currently unknown. Future work in the CHORDS system could include a survey of all 14 CHORDS partners to understand whether more granular race, ethnicity, or national origin data are collected. If enough partners are collecting these data, CHORDS could plan to load these fields into the VDW so its data experts can assess the characteristics of patients who have more granular race, ethnicity, or national origin data.

CHORDS also compiles data on patients' primary language spoken during an encounter. The language field adheres to the ISO-639-2 specifications for language categories. CHORDS data experts are also interested in exploring the potential use of primary language data as a proxy for national origin.

All CHORDS partners have some missing race/ethnicity data for their patient population. CHORDS experts are interested in examining the number of patients missing race/ethnicity data in 2019 and 2020 and could investigate approaches for multiple imputation for a distributed data network. This process would include: requesting a masked sample EHR dataset to test a raking method that has previously been used on EHR data to understand its effectiveness; examining whether PopMedNet can be used to impute data; and evaluating the benefits and challenges of imputing data from patients with known race/ethnicity data within the same organization and across organizations. Because CHORDS is a distributed data network, EHRs from other systems might be used to predict a patient's race/ethnicity.

Emily Bacon, PhD, Bacon Analytics LLC, contributed to this section.

The Colorado Health Access Survey

Description

The Colorado Health Access Survey (CHAS) is Colorado's premier source of data on health coverage, access to care, and affordability.[75] 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.

State of Race/Ethnicity Data

Until 2021, the CHAS survey incorporated two questions: one on Hispanic/Latino ethnicity and one on race, using the major choices of white, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, or Some Other Race. Hispanic was also asked as a racial category when a respondent indicated Hispanic ethnicity. See Figure 4.

Figure 4. CHAS Race and Ethnicity Questions, 2009-2019*

1. Are you Hispanic or Latino? • Yes • No, not of Hispanic origin • Don't know • Refused/Blank
2. Which one or more of the following would you say is your race? You may select more than one. • White • Black or African American • Asian • Native Hawaiian or other Pacific Islander • American Indian or Alaska Native • Hispanic (*asked only if respondent answers Yes to Item 1.*) • Some other race (specify) • Don't know • Refused/Blank

* Survey instructions have been simplified for display purposes.

Several changes were made to the 2021 survey that aligned with CHI's exploration of racial equity issues and consultation with stakeholders. Among the changes were adding ethnic

identity questions, including a Middle Eastern or North African racial category, and alphabetizing the race categories. The response categories are displayed in Figure 5 in Prospective Developments. These response categories were based on other existing questionnaires and on feedback from stakeholders. The feedback from external stakeholders, which is described in greater detail in the Lessons Learned section of this report, was particularly useful in supporting these changes.

Limitations

The CHAS has its own set of limitations with data disaggregation. A survey of about 10,000 people (out of 5.8 million Coloradans) may not be powerful enough to identify a representative sample of the relatively small racial and ethnic communities in the state, despite oversampling strategies. The data are all self-reported with no way to verify responses.

Respondents on the survey remain anonymous so CHI does not have identifiers such as Social Security Number or birth date to undertake direct matching with other data sources.

Changes in methods to collect race/ethnicity may result in the inability to combine years of data and trend the data over time. CHI is exploring the bridging methods described in this report to model how people who identify as Middle Eastern or North African, for example, may have responded in past surveys.

Despite its limitations, the CHAS remains a rich source of information on the health of Coloradans. It is the only source for some types of information, such as detailed access to care questions and circumstances of why some Coloradans lack health coverage.

Prospective Developments

As described in this report's Colorado use case, the CHAS will serve as the primary source of disaggregated data on the Hispanic/Latinx community for this project. Additional fields collected on the CHAS are outlined below in Figure 5.

Figure 5. CHAS Race and Ethnic Identity Questions, 2021*

1. Are you Hispanic or Latino?
    - Yes
    - No, not of Hispanic origin
    - Don't know
    - Refused/Blank

(Asked of people who respond "Yes" to Item 1.)

1a. Please indicate how you identify or represent yourself. (*Mark all that apply.*)
- Mexican/Mexican American
- Chicano
- Central American (El Salvador, Guatemala, Honduras, Nicaragua, Panama, etc.)
- South American (Chile, Colombia, Ecuador, Peru, Venezuela, etc.)
- Caribbean (Cuba, Dominican Republic)
- Latinx[76]
- Spanish-American (from Spain)
- Something else (Specify: _____)
- Don't know

- Refused/Blank

2. Which one or more of the following would you use to describe yourself? Would you describe yourself as… (*Mark all that apply.*)
   - American Indian or Alaska Native
   - Asian
   - Black or African American
   - Hispanic/Latino (*asked only if respondent answers Yes to Item 1.*)
   - Middle Eastern or North Africa
   - Native Hawaiian or Other Pacific Islander
   - White
   - Some other race (Specify: _____)
   - Don't know
   - Refused/Blank

(Asked of people who respond "Asian" to Item 2.)

3. You said Asian. Which group best represents your heritage or ancestry? (*Mark all that apply.*)

   - Bangladeshi
   - Burmese
   - Cambodian
   - Chinese
   - Filipino
   - Hmong
   - Indian (India)
   - Indonesian
   - Japanese
   - Korean
   - Laotian
   - Malaysian
   - Pakistani
   - Sri Lankan
   - Taiwanese
   - Thai
   - Vietnamese
   - Something else? (Please specify what other group best represents your Asian heritage or ancestry:_____)
   - Don't know
   - Refused/Blank

(Asked of people who respond "Native Hawaiian or Other Pacific Islander" to Item 2.)

4. You said you were Native Hawaiian or Other Pacific Islander. Which group best represents your heritage or ancestry? (*Mark all that apply.*)
   - Native Hawaiian
   - Guamanian or Chamorro
   - Samoan
   - Something else? (Please specify what other group best represents your Pacific Islander heritage or ancestry:_____)
   - Don't know
   - Refused/Blank

(Asked of people who respond "Black or African American" to Item 2.)

5. You said Black or African American. Which group best represents your Black heritage or ancestry? (*Mark all that apply.*)
    - African American
    - Caribbean or West Indian
    - A recent immigrant or the child of recent immigrants from Africa
    - Something else? (Please specify what other group best represents your Black or African American heritage or ancestry:_____)
    - Don't know
    - Refused/Blank

(Asked of people who respond "Middle Eastern or North African" to Item 2.)

6. You said Middle Eastern or North African. Which group best represents your Middle Eastern or North African heritage or ancestry? (*Mark all that apply.*)
    - Algeria
    - Bahrain
    - Djibouti
    - Egypt
    - Gaza
    - Iran
    - Iraq
    - Israel
    - Jordan
    - Kuwait
    - Lebanon
    - Libya
    - Mauritania
    - Morocco
    - Oman
    - Qatar
    - Saudi Arabia
    - Sudan
    - Syria
    - Tunisia
    - United Arab Emirates
    - West Bank
    - Yemen
    - Something else? (Please specify what other group best represents your Middle Eastern or North African heritage or ancestry:_____)
    - Don't know
    - Refused/Blank

* Survey instructions have been simplified for display purposes.

References

[1] U.S. Census Bureau. (2020). 2019 American Community Survey, 1 Year Estimates. Retrieved 5 Feb 2021 from https://data.census.gov/cedsci/table?q=race&g=0400000US08&tid=ACSDP1Y2019.DP05&hidePreview=true.

[2] Wang, H. (2018). No Middle Eastern Or North African Category On 2020 Census, Bureau Says. NPR. https://www.npr.org/2018/01/29/581541111/no-middle-eastern-or-north-african-category-on-2020-census-bureau-says.

[3] Wang, H. (2020). Biden Wants Census to See 'Invisible' Groups: LGBTQ, Middle Eastern, North African. NPR. https://www.npr.org/2020/11/14/932594879/biden-wants-census-to-see-invisible-groups-lgbtq-middle-eastern-north-african#:~:text=No%20Middle%20Eastern%20Or%20North,their%20own%20on%20census%20forms.

[4] Harron, K., Goldstein, H., and Dibben, C. (2016). Methodological Developments in Data Linkage. Wiley.

[5] The National Academy of Sciences, Engineering, and Medicine. (2017). Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. The National Academies Press. https://doi.org/10.17226/24893.

[6] Donatiello, G., et al. (2014). Statistical Matching of Income and Consumption Expenditures. International Journal of Economic Sciences, 3(3), 50–65.

[7] Johnson, J., et al. (2009). Tribal Linkage and Race Data Quality for American Indians in a State Cancer Registry. Am J Prev Med, 36(6), 549–554. DOI: 10.1016/j.amepre.2009.01.035.

[8] Kum, H. and Masterson, T. (2008). Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being. Levy Economics Institute of Bard College. http://www.levyinstitute.org/publications/statistical-matching-using-propensity-scores.

[9] Donatiello, G., et al. (2014).

[10] Xue, Y., Harel, O., and Aseltine Jr., R. (2019). Imputing race and ethnic information in administrative health data. Health Services Research, 54, 957–963. DOI: 10.1111/1475-6773.13171.

[11] Brown, D., Knapp, C., Baker, K., and Kaufmann, M. (2016). Using Bayesian Imputation to Assess Racial and Ethnic Disparities in Pediatric Performance Measures. 51(3), 1095–1108.

[12] Schenker, N., and Parker, J. (2003). From single-race reporting to multiple-race reporting: using imputation methods to bridge the transition. Statistics in Medicine, 22, 1571-1587.

[13] Elliot, M., et al. (2008). A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. Health Services Research, 45(5), 1722–1736.

[14] Adjaye-Gbewonyo, D., et al. (2014). Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study. Health Services Research, 49(1), 268–283.

[15] Coley, R., Johnson., E, and Simon, G. (2021). Racial/Ethnic Disparities in Performance of Prediction Models for Death by Suicide After Mental Health Visits. JAMA Psychiatry. https://jamanetwork.com/journals/jamapsychiatry/article-abstract/2778923

[16] Vatsalan, D., Sehili, Z., Christen, P., and Rahm, E. (2017). Privacy-Preserving Record Linkage for Big Data. Current Approaches and Research Challenges. In: Zomaya, A., Sakr, S. (eds) Handbook of Big Data Technologies. 851–895. Springer. DOI: 10.1007/978-3-319-49340-4_25.

[17] Bowen, C. Williams, A. R., and Narayanan A. (2021). To Advance Racial Equity, Releasing Disaggregated Data while Protecting Privacy Will Be Key. Urban Institute. https://www.urban.org/urban-wire/advance-racial-equity-releasing-disaggregated-data-while-protecting-privacy-will-be-key.

[18] Center for Improving Value in Health Care. (2021). About CIVHC. Retrieved from https://www.civhc.org/.

[19] The Colorado Health Observation Regional Data Service. (2021). CHORDS. Colorado Health Institute. https://www.coloradohealthinstitute.org/research/CHORDS.

[20] van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. https://www.jstatsoft.org/v45/i03/.

[21] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.
[22] University of Michigan Institute for Social Research. (2021). IVEware: Imputation and Variance Estimation Software. Retrieved from https://www.src.isr.umich.edu/software/iveware/.
[23] Thompson, C., Boothroyd, D., Hastings, K., Cullen, M., Palaniappan, L., and Rehkopf, D. (2018). A Multiple-Imputation 'Forward Bridging' Approach to Address Changes in the Classification of Asian Race/Ethnicity on the US Death Certificate. American Journal of Epidemiology, 187(2), 347-357.
[24] Schenker, N. and Parker, J. (2003).
[25] Ingram, D. D., Parker, J. D., Schenker, N., Weed, J. A., Hamilton, B., Arias, E., Madans J. H. (2003). United States Census 2000 population with bridged race categories. National Center for Health Statistics. Vital Health Stat, 2(135).
[26] Zou, K, O'Malley, J, and Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation, 115(5): 654-657.
[27] U.S. Department of Health and Human Services. (2002). Studies of Welfare Populations: Data Collection and Research Issues. Two Methods of Linking: Probabilistic and Deterministic Record-Linkage Methods. https://aspe.hhs.gov/report/studies-welfare-populations-data-collection-and-research-issues/two-methods-linking-probabilistic-and-deterministic-record-linkage-methods.
[28] Harron, K., Goldstein H., and Dibben, C. (2016).
[29] Harron, K., Goldstein H., and Dibben, C. (2016).
[30] National Survey for Health Statistics. (2012). Linkages Between Survey Data from the National Center for Health Statistics and Medicare Program Data from the Centers for Medicare and Medicaid Services. https://www.cdc.gov/nchs/data/datalinkage/cms_medicare_methods_report_final.pdf.
[31] U.S. Department of Health and Human Services. (2002).
[32] Harron, K., Goldstein H., and Dibben, C. (2016).
[33] Shlomo, N. (2019). Overview of Data Linkage Methods for Policy Design and Evaluation. Data-Driven Policy Impact Evaluation. Springer. https://doi.org/10.1007/978-3-319-78461-8_4.
[34] Shlomo, N. (2019).
[35] Harron, K., Goldstein H., and Dibben, C. (2016).
[36] Kumar, R. G., et al. (2018). Probabilistic Matching of Deidentified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center: A Follow-up Validation Study. Am J Phys Med Rehabil. 97(4), 236-241. doi:10.1097/PHM.0000000000000838.
[37] Kesinger, M., et al. (2017). A Probabilistic Matching Approach to Link De-identified Data from a Trauma Registry and a Traumatic Brain Injury Model System Center. Am J Phys Med Rehabil. 96(1): 17–24. doi:10.1097/PHM.0000000000000513.
[38] Leulescu, A., and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat. https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-RA-13-020.
[39] Singh, A., Mantel, H., Kinack, M., and Rowe, G. (1993). "Statistical Matching: Use of Auxiliary Information as an Alternative to the Condition Independence Assumption." Survey Methodology, 19 (1): 59-79.
[40] Huber, M., and Melly, B. (2012). A Test of the conditional independence assumption in sample selection models. http://dx.doi.org/10.2139/ssrn.2151357.
[41] Singh, A., Mantel, H., Kinack, M., and Rowe, G. (1993).
[42] Leulescu, A., and Agafitei, M. (2013).
[43] Leulescu, A and Agafitei, M. (2013).
[44] Leulescu, A and Agafitei, M. (2013).
[45] Sterne, J. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. https://www.bmj.com/content/338/bmj.b2393.
[46] Harron, K, Goldstein H, and Dibben, C. (2016).
[47] Sterne, J. (2009).
[48] Sterne, J. (2009).

[49] Zhang, G., Parker, J., and Schenker, N. (2016). Multiple Imputation for Missingness Due to Nonlinkage and Program Characteristics: A Case Study of the National Health Interview Survey Linked to Medicare Claims. J Surv Stat Methodol, 4(3), 316–338. DOI:10.1093/jssam/smw002.
[50] Silva, G., Trivedi, A., and Gutman, R. (2019). Developing and evaluating methods to impute race/ethnicity in an incomplete dataset. Health Serviced and Outcomes Research Methodology, 19: 175–195.
[51] Silva, G, Triveldi, A, and Gutman, R. (2019). 1–21.
[52] Parker, J., Schenker, N., Ingram, D., Weed, J., Heck, K., and Madans, J. (2004). Bridging Between Two Standards for Collecting Information on Race and Ethnicity: An Application to Census 2000 and Vital Rates. Public Health Reports, 119, 195-205.
[53] Schenker, N., and Parker, J. (2003).
[54] Schenker, N., and Parker, J. (2003).
[55] Parker, J., Schenker, N., Ingram, D., Weed, J., Heck, K., and Madans, J. (2004).
[56] Thompson, C., Boothroyd, D., Hastings, K., Cullen, M., Palaniappan, L., and Rehkopf, D. (2018).
[57] Bender, R. (2009) Introduction to the Use of Regression Models in Epidemiology. In: Verma M. (eds) Cancer Epidemiology. Methods in Molecular Biology, vol 471. Humana Press. https://doi.org/10.1007/978-1-59745-416-2_9
[58] Sullivan, L., and LaMorte, Q. (n.d.) Multivariable Methods. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/BS704-EP713_MultivariableMethods_print.html.
[59] Pennsylvania State University. (2018). 10.4 – Multicollinearity. https://online.stat.psu.edu/stat462/node/177/.
[60] Pennsylvania State University. (2018). 10.7 – Detecting Multicollinearity Using Variance Inflation Factors. https://online.stat.psu.edu/stat462/node/180/.
[61] Whaley, A. (2003). Ethnicity/race, ethics, and epidemiology. Journal of the National Medical Association, 95(8), 736–742.
[62] Elkind, M., et al. (2020). Approaches to Studying Determinants of Racial-Ethnic Disparities in Stroke and Its Sequelae. Stroke, 51(11), 3406–3416.
[63] Nipp, R., et al. (2018). Disparities in cancer outcomes across age, sex, and race/ethnicity among patients with pancreatic cancer. Cancer Med, 7(2), 525–535.
[64] Elliot, M et al. (2008).
[65] Adjaye-Gbewonyo, D et al. (2014).
[66] American Psychological Association. (2017). Ethnic and Racial Minorities and Socioeconomic Status. https://www.apa.org/pi/ses/resources/publications/minorities.
[67] Passel, J., and Cohn, D. (2019). Mexicans decline to less than half the U.S. unauthorized immigrant population for the first time. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/06/12/us-unauthorized-immigrant-population-2017/.
[68] Center for Improving Value in Health Care. (2021).
[69] Center for Improving Value in Health Care. (2021). Colorado APCD Race and Ethnicity Data. https://www.civhc.org/get-data/publications/.
[70] Center for Improving Value in Health Care. (2021). CO APCD Insights: Methodological Notes. https://www.civhc.org/wp-content/uploads/2021/02/COAPCD_InsightsDashboard_Methodology_Final-2.2021.pdf.
[71] Scott, K., et al. (2020). Evaluating Population Coverage in a Regional Distributed Data Network: Implications for Electronic Health Record–Based Public Health Surveillance. Public Health Reports, 135(5), 621–630.
[72] The Colorado Health Observation Regional Data Service. (2021).
[73] The Colorado Health Observation Regional Data Service. (2021).
[74] The Colorado Health Observation Regional Data Service. (2021). 2019 Care Population [Data set]. https://www.coloradohealthinstitute.org/research/CHORDS.
[75] Colorado Health Institute. (2019). 2019 Colorado Health Access Survey: Progress in Peril. https://www.coloradohealthinstitute.org/research/CHAS.
[76] While CHI does not consider "Latinx" to be an ethnicity, key stakeholders advised adding this category to assess its acceptance within Colorado's Latinx/Hispanic communities.