

**Analytic Data Framework**

Contents

[1. List of Abbreviations 3](#_Toc106787362)

[2. Background 3](#_Toc106787363)

[3. Overarching Goals 4](#_Toc106787364)

[4. Specific Aims 4](#_Toc106787365)

[5. Populations 4](#_Toc106787366)

[6. Exposure Groups 5](#_Toc106787367)

[7. CDE Variables 5](#_Toc106787368)

[8. Other Existing Internal Sources 10](#_Toc106787369)

[9. External Data Sources 10](#_Toc106787370)

[10. Data Linkage 11](#_Toc106787371)

[11. Handling of Missing Data 11](#_Toc106787372)

[12. Core Analytic Datasets 13](#_Toc106787373)

[13. Core Analytic Dataset Data Dictionary 14](#_Toc106787374)

[14. SAP Template 14](#_Toc106787375)

[15. Other CDCC Resources 14](#_Toc106787376)

# List of Abbreviations

|  |  |
| --- | --- |
| ACS | American Community Survey |
| ADI | Area Deprivation Index |
| AHRF | Area Health Resource Files |
| AHRQ | Agency for Healthcare Research and Quality  |
| AIDS | Acquired Immunodeficiency Syndrome |
| BRFSS | Behavioral Risk Factor Surveillance System |
| CBSA | Core-Based Statistical Area |
| CDC | Centers for Disease Control and Prevention |
| CDCC | Coordination and Data Collection Center |
| CDE | Common Data Element |
| COI | Child Opportunity Index |
| COPD | Chronic Obstructive Pulmonary Disease |
| COVID-19 | Coronavirus Disease 2019 |
| CPR | Community Profile Report |
| EHR | Electronic Heath Records |
| HIFLD | Homeland Infrastructure Foundation-Level Data |
| HIV | Human Immunodeficiency Virus |
| HRSA | Health Resources and Services Administration |
| LILA | Low-Income and Low-Access |
| NIH | National Institute of Health |
| NYU | New York University |
| PPE | Personal Protective Equipment |
| PVI | Pandemic Vulnerability Index |
| RADx-UP | Rapid Acceleration of Diagnostics for Underserved Populations |
| RUCA | Rural-Urban Commuting Area |
| SAP  | Statistical Analysis Plan |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome Coronavirus 2 |
| SDI | Social Deprivation Index |
| SES | Socioeconomic Status |
| SVI | Social Vulnerability Index |
| SYCT  | Say Yes! COVID Test |
| USDA | U.S. Department of Agriculture |

# Background

The Rapid Acceleration of Diagnostics for Underserved Populations (RADx-UP) program includes a Coordination and Data Collection Center (CDCC) as well as >100 community-engaged projects to evaluate strategies to increase access and uptake of SARS-CoV-2 tests in underserved populations across the United States. This network of projects also includes the collection and analysis of data on disparities in infection rates, disease progression and outcomes.

This Analytic Data Framework addresses the overarching goals specified for the RADx-UP program under the responsibility of the CDCC by describing the data required for analysis to meet program goals, data being collected by projects, datasets to be linked, discrepancies between the data required and data expected, and data specifications for analytic datasets to be distributed by the CDCC. Separate statistical analysis plans (SAP) will be developed for cross-consortium data analysis objectives to meet program-level aims.

# Overarching Goals

The overarching goals of the RADx-UP initiative are to:

* Reduce disparities in COVID-19 associated morbidity and mortality
* Lay the foundation to reduce disparities for underserved populations

# Specific Aims

These RADx-UP program goals will be accomplished by *strengthening the available data* on:

* Disparities in infection rates, disease progression, and outcomes
* Differences in testing access and uptake patterns
* Identifying strategies to address disparities in COVID-19 testing

# Populations

Analytic populations will consist of participants enrolled in RADx-UP projects, which broadly includes populations that are *underserved* as well as populations that are *COVID-19 vulnerable* due to medical, geographic and social factors. Data on other populations (neither underserved nor COVID-19 vulnerable) may be included based on analytic needs.

Underserved populations:

* + - * American Indians/Alaska Natives
			* Asian Americans
			* Blacks/African Americans
			* Hispanics/Latinos
			* Native Hawaiians/ other Pacific Islanders
			* Sexual and gender minorities
			* Socioeconomically disadvantaged populations
			* Underserved rural populations

COVID-19 medically and/or socially vulnerable populations:

* + - * Residents of nursing homes and assisted living facilities
			* Community-dwelling older adults
			* Individuals with intellectual, developmental, sensory, or physical disabilities
			* Individuals with cognitive impairment or dementia, or communication disorders
			* Homeless populations
			* Individuals involved with the criminal or juvenile justice systems (incarcerated or under community supervision)
			* Individuals with medical comorbidities known to increase risk of severe COVID-19, including heart failure and related cardiovascular conditions, diabetes mellitus, chronic lung disease, obesity, HIV/AIDS; pregnant and post-partum women
			* Pregnant Women
			* Children and adolescents
			* Individuals living in congregate housing such as shelters or residential treatment facilities
			* Individuals in overcrowded housing
			* Individuals with substance use disorders or serious mental illness
			* Migrant and immigrant populations
			* Residents of tribal lands or reservations
			* Communities exposed to high rates of air pollution or other toxic exposures
			* Rural and remote communities

# Exposure Groups

The exposure groups are subgroups of populations in which metrics may be reported. The exposure groups may vary based on the RADx-UP program needs. Below is a list of potential exposure groups that can be used in the analysis. This assumes sample sizes at the given level of granularity are large enough to support the analysis.

* Racial minority
* Sexual orientation and gender minorities
* Socioeconomically disadvantaged populations
* Underserved rural populations
* Ethnicity (Hispanic and the subpopulations of place of origin [Cuban, Puerto rican, etc])
* People who are incarcerated
* Older adults (65+)
* Children
* Pregnant women
* People in nursing homes
* People with comorbidities at increased risk of COVID-19
* Non-English speakers
* Rural/Urban populations
* People who are Insured/Uninsured
* Frontline essential workers
* People with disabilities
* Unemployed population
* People with food insecurity
1. CDE Variables

The CDCC is charged with managing collection, integration, and sharing of research projects data for the RADx-UP consortium. This section provides a summary of CDE study data collected by the CDCC in efforts to standardize, harmonize and integrate for cross-consortium analysis.

EHR data can be mapped to most CDE elements, and those in the CDE will be included for analysis. Mapping guidance can be found within the radx-up link (<https://radx-up.org/wp-content/uploads/2021/10/RADx-UP-EHR-Guidance-2021_07_26.pdf>). CDE versions, codebook and additional information can be found in the NIH RADx-UP Common Data Elements link ([radx-up.org/research/cdes/](https://radx-up.org/research/cdes/)).

|  |  |
| --- | --- |
| **CDE - Tier 1** | **CDE - Tier 2** |
| **Location** |  |
| * County
* Zip Code
 |  |
| **Sociodemographics** |  |
| * Race
 | * Sexual orientation (additional description)
 |
| * + Asian origin
 |  |
| * + Native Hawaiian or Other Pacific Islander origin
 |  |
| * Ethnicity
* Hispanic detail
 |  |
| * Age
 |  |
| * Sex at birth
* Sexual orientation
 |  |
| * Gender Identity
 |  |
| * Pregnancy status
 |  |
| * Years of education
 |  |
| **Housing / Employment / Insurance / Transportation** |
| * Housing composition
	+ Homelessness
* Congregate Housing
 | * Staying in the same place the past two months
* Concerned about not having a place to live in the next two months
 |
| * Loss of employment income
 |  |
| * Current employment status
	+ Essential worker (Yes/No)
	+ Job type
 |   |
| * Health insurance coverage type
 |   |
| * Health coverage loss due to COVID
 |   |
| * Getting health care needed
 |   |
| * Place to stay/live
 |   |
| * Getting enough food
 |   |
| * Having clean water to drink
 |   |
| * Getting medicine needed
 |   |
| * Getting to where needed to go
 |   |
| * Languages spoken at home
* Preferred language spoken at home
 |  |
| * Family income
 |   |
| **Work PPE and Distancing** |  |
| * Access to necessary washing facilities
* Work requires close contact
* Access to necessary PPE at work
 |   |
| **Medical History** |  |
| * Immunocompromised condition Autoimmune disease
* Hypertension
* Diabetes
* Chronic kidney disease
* Cancer
* Cardiovascular disease
* Asthma
* COPD
* Other chronic lung disease
* Sickle cell anemia
* Depression
* Alcohol or substance use disorder
* Intravenous drug use
* Other mental health disorder
* Other chronic condition
 | * Needed to postpone any medical care since start of COVID-19 pandemic (March 2020)
 |
| **Health Status** |  |
| * Height
	+ Feet and inches
	+ Meters and centimeters
* Weight
	+ Kilograms
	+ Pounds
* Self-reported health
* Self-reported disability
 |  |
| **Vaccine Acceptance** |  |
| * Ever received flu vaccine
	+ Received flu vaccine in last 12 months
* Received COVID-19 vaccine
	+ Likelihood to get COVID-19 vaccine when available
* Reasons to get COVID-19 vaccine
* Concerns about getting COVID vaccine
 | * The vaccine is safe
* The vaccine prevents COVID-19
* The vaccine is free or low cost
* The vaccine is not painful
* The vaccine is convenient (where/when)
* Vaccine manufacturer
* Number of vaccine doses received
	+ Date of first vaccine shot
	+ Date of most recent vaccine shot
* Given information or brochures about vaccine and/or COVID-19 before vaccination
* Given information or brochures about vaccine and/or COVID-19 after vaccination
* How often needed help to understand information or guidance from doctor, nurse or pharmacist
* Known prior exposure to someone who tested positive for COVID-19
* Informed about which vaccine option is right for you
* Vaccine required by school or job
* Intent to get vaccine booster if recommended
 |
| **Testing** |  |
| * Ability to isolate without losing job
 | * Confidence in negative test result
 |
| * Ability to quarantine without losing job
 | * Confidence in positive test result
 |
| * Ever been tested
	+ Ever tested positive
 | * Testing encouraged from reduced worry that I might have COVID-19
 |
| * + - Time first tested positive (mm/yyyy)
* Time of most recent test (mm/yyyy)
 | * Testing encouraged from believing that I was exposed to someone who has COVID-19
 |
| * + Result of most recent test
	+ Most recent test method
 | * Testing encouraged to know if I am safe not to give COVID-19 to friends and family
 |
| * Knowledge of COVID testing access
* Ease of access to COVID testing
 | * Testing encouraged to know if I am safe not to give COVID-19 to anyone I am around
 |
|  | * Testing encouraged to let my employer know that I am safe to work
 |
|  | * Testing encouraged to get treated early (if positive)
 |
|  | * Testing discouraged from discomfort
 |
|  | * Testing discouraged from even if I don't have it when tested, I can still get COVID-19 later.
 |
|  | * Testing discouraged from no COVID-19 symptoms so I don't need to be tested
 |
|  | * Testing discouraged from if positive, officials will need to contact the people I've been in contact with
 |
|  | * Testing discouraged from not wanting to know if I have it
 |
|  | * Plan to get tested as often as needed
 |
|  | * Interpretation of negative test result
 |
|  | * Interpretation of positive test result
 |
|  | * Anyone close become sick from COVID
 |
|  | * Anyone close been hospitalized from COVID
 |
|  | * Anyone close died from COVID
 |
|  | * Had COVID
 |
|  | * Personal level of risk for getting sick from COVID-19
 |
|  | * Tested for COVID-19 in the last 30 days
 |
|  | * Primary reason for latest test
 |
|  | * Barriers or problems encountered with testing
 |
|  | * Employer offer paid time off if positive test
 |
|  | * Challenges to isolate due to a positive test or illness
 |
|  | * Wait time between scheduling the test and when the test was performed
 |
|  | * How are test results communicated
 |
| **COVID Testing** |  |
| * Date of COVID test
 |   |
| * COVID testing participant disease status
 |   |
| * COVID test approval
 |   |
| * COVID test collection setting
 |   |
| * COVID test location performed
 |   |
| * COVID test study setting
 |   |
| * COVID test type
 |   |
| * COVID test specimen type
 |   |
| * COVID test specimen collector
 |   |
| * Date and time specimen collected
 |   |
| * Date and time result
 |   |
| * COVID test result
 |   |
| **Symptoms** |  |
| * Symptom date
 |   |
| * Symptoms past week
 |   |
| * Fever or chills
 |   |
| * Cough
 |   |
| * Shortness of breath or difficulty breathing
 |   |
| * Lack of energy or general tired feeling
 |   |
| * Muscle or body aches
 |   |
| * Headache
 |   |
| * New loss of taste or smell
 |   |
| * Sore throat, congestion or runny nose
 |   |
| * Feeling sick to your stomach or vomiting, diarrhea
 |   |
| * Abdominal Pain
 |   |
| * Skin Rash
 |   |
| * Other
 |   |
| **Medications** |  |
|  | * Currently take prescription medications
	+ Prescription medication name(s)
 |
| **Alcohol and Tobacco** |  |
| * Ever drank alcohol
	+ How often drink alcohol
* Current smoking status
	+ Cigarettes per day
* Current vaping status
 | * Number of alcoholic drinks on a typical drinking day
* Years of smoking
* Years of vaping
 |
| **Drug Use** |  |
|  | * Marijuana use in past 12 months
	+ Marijuana smoking frequency
	+ Marijuana vaping frequency
 |
|  | * Prescription drug use just for the feeling, more than prescribed, or not prescribed for you in past 12 months
 |
|  | * Used cocaine or crack, heroin, crystal meth, hallucinogens, ecstasy in past 12 months
 |
|  | * + Cocaine or crack use frequency
	+ Heroin use frequency
	+ Crystal meth use frequency
	+ Hallucinogen use frequency

Ecstasy use frequency |
| **Disability** |  |
|  | * Deaf or serious hearing difficulty
* Blind or serious vision difficulty, even when wearing glasses
* Serious difficulty concentrating, remembering, or making decisions due to physical, mental, or emotional condition
* Serious difficulty walking or climbing stairs
* Difficulty dressing or bathing
* Difficulty doing errands alone, such as visiting a doctor office or shopping, due to physical, mental, or emotional condition
 |
| **Food Insecurity** |  |
|  | * Food bought didn’t last and didn’t have money to get more
* Couldn’t afford to eat balanced meals
* Cut size of or skipped meals due of lack of money for food
	+ Frequency of any of these three
	+ Eat less than felt you should due to lack of money to buy food
 |
|  | * Ever hungry but didn’t eat because couldn’t afford enough food
 |
|  | * High quality fresh fruits and vegetables in neighborhood
 |
|  | * Doctor or health care provider
* Faith leader
* Close friends and family members
* Colleagues, classmates, acquaintances
* News on radio, TV, online, or in newspapers
* Social media contacts
* U.S. government
* U.S. Coronavirus Task Force
 |

# Other Existing Internal Sources

Additional internal data sources are administrative, operational and study data collected by the CDCC not listed in the previous section.

* Additional CDE metrics data (e.g. PhenX, BRFSS, etc) not mappable to NIH RADx-UP CDEs
* Qualitative data
* Operational data in REDCap
* Project management data in Asana
* Research Performance Progress Report
* Evidence Academy – Exit Survey
* Evidence Academy – Registration Data
* Evidence Academy – Focus Groups/Interviews
* Testing Assay Database
* Project Feedback Loop
* Project Requests
* Intake Survey
* Core-specific report, tracking spreadsheet, or information request
* Pilot project survey
* Publications tracking database
* Altmetrics
* Web analytics
* Biosketches
* Working Group Attendance Logs
* Working Group Activation Plans
* Working Group Feedback Surveys
* Network Survey
* RADx-UP Project Implementation Survey (Track and Evaluation team)
* CDCC and Partners’ Survey (Track and Evaluation team)
* RADx-UP Project Feedback Survey (Track and Evaluation team)
* Core Routine Monitoring Survey (Track and Evaluation team)

# External Data Sources

The CDCC is also responsible with linking the CDE data to publically available sources. This section provides a summary of external data sources that are currently linked and available to the CDCC.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data**  | **Source** | **Data Description** | **Identifier used for Linkage** | **Source Location/Data Dictionary** |
| Social Vulnerability Index (SVI) | CDC/ ATSDR | SVI ranks each census tract on 15 U.S. census social factors, including poverty, lack of vehicle access, and crowded housing, and groups them into four related themes – socioeconomic, household, race, housing/transportation to help local officials identify communities that may need support before, during, or after disasters. | Zip code | [https://www.atsdr.cdc.gov/ placeandhealth/svi/index.html](https://www.atsdr.cdc.gov/%20placeandhealth/svi/index.html) |
| Area Deprivation Index (ADI) | University of Wisconsin | Composite index based on socioeconomic indicators of neighborhood deprivation, such as income, education, employment, and housing quality. It is presented in ranking deciles and percentiles. The ADI also allows for the ranking of neighborhoods at the Census Block Group level by socioeconomic disadvantage and can be used to inform health delivery and policy, especially for the most disadvantaged neighborhood groups. | Zip code | https://www.neighborhoodatlas. medicine.wisc.edu/  |
| AHRQ Socioeconomic Status (SES) | AHRQ | A weighted measure of multiple indicators of neighborhood deprivation that includes racial and socioeconomic status. Data provided at census tract level. | Zip code | [https://archive.ahrq.gov/research /findings/final-reports/medicareindicators/ index.html#contents](https://archive.ahrq.gov/research%20/findings/final-reports/medicareindicators/) |
| Child Opportunity Index (COI) | Brandeis University | The COI measures and maps the quality of resources and conditions that matter for children to develop in a healthy way in the neighborhoods where they live. The COI 2.0 is a composite index measured at the census tract level that captures neighborhood resources and conditions spanning 29 indicators and three domains (education, health and environment, and social and economic). | Zip code | https://www.diversitydatakids.org /child-opportunity-index |
| American Community Survey (ACS)  | US Census Bureau | Geographic level data on social, economic, housing, and demographic characteristics. This is the 2015-2019 5-year estimates data. | Zip code | <https://www.census.gov/programs-surveys/acs/data.html>  |
| Pandemic Vulnerability Index (PVI)  |  NIH | The PVI summarizes and visualizes overall risk in a radar chart, where different data sources, such as infection rates, population mobility, demographics, air pollution, make up pieces of the pie. |  County | [https://www.niehs.nih.gov/research/](https://www.niehs.nih.gov/research/programs/coronavirus/covid19pvi/%20details/index.cfm)[programs/coronavirus/covid19pvi/ details/index.cfm](https://www.niehs.nih.gov/research/programs/coronavirus/covid19pvi/%20details/index.cfm) |
| Community Profile Report (CPR) COVID-19 positivity rate | CDC | A snapshot in time of COVID-19 positivity rates in the last seven days.  | County | <https://healthdata.gov/Health/COVID-19-Community-Profile-Report/gqxm-d9w9> |
| CDC COVID-19 Vaccination Coverage  | CDC | Vaccination coverage data.  | County | [https://data.cdc.gov/Vaccinations/ COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh](https://data.cdc.gov/Vaccinations/%20COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh)  |

# Data Linkage

The strategy for data linkage will vary according to the granularity of geographic data available for data being linked.

* External data sets that are available at the Zip code level will be linked to data at the CDCC based on Zip code where Zip code is available.
* External data sets that are available at the county level will be linked to data at the CDCC based on county where county is available.
* No geographic information: see Section 11, Handling of Missing Data

# Handling of Missing Data

The handling of missing data will follow the workflow below, which will guide imputation strategies, where warranted, based on the degree, type and mechanism of missingness. These same guidelines will be applied to all datasets modified for specific analysis.

****

**Handling of Missing Data Examples**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Objective** | **Data Source** | **Type of Missing Data** | **Degree and Mechanism of Missingness Assessment** | **Handling of Missing Data** | **Imputation Strategy** |
| **Project-level Missingness****Example** | Describe perceived COVID-19 testing site accessibility by levels of Housing Type & Transportation Social Vulnerability Index | CDESVI (Housing type & transportation) | Project-level missing data. One small project did not collect Zip code from participants.  | Five percent of participants have missing Zip codes. Missing participants’ characteristics are homogenous, and data collection was focused on a specific area of a city. Geographic distribution crucial for analysis. | Use the project’s targeted area Zip code for all participants in the project as proxy | Imputation from external variable |
| **Participant-level** **Missingness****Example** | Estimate number of COVID-19 tests performed by socioeconomic status (AHRQ SES) | CDEAHRQ SES | Participant-level missing data in linkage. Participants did not consent to release Zip code.  | One quarter of participants have missing Zip codes and are all Hispanic, but all other socioeconomic variables are available. Geographic distribution not crucial for analysis. | Use CDE socioeconomic variables to impute AHRQ SES values for all participants with missing Zip code | Participant-level multiple imputation from non-linked variable |

# Core Analytic Datasets

A series of analytic datasets will be created to support cross-consortium analyses addressing the objectives described in this document. Specifications and data dictionaries will be developed and provided in the Appendix. These Core Analytic Datasets will be organized into domains and three groups:

1. Participant-level data: participant-level CDE data collected at one point in time, where one row represents one participant, e.g. demographics, medical history, etc.
2. Record-level data: CDE data of multiple records per participant where one row represents one record, e.g. longitudinal data.
3. Linked data: participant-level CDE data that contains variables from external data sources through linkage.

The first phase of these Core Analytic Datasets will include Tier 1 CDEs, and the second phase will include some Tier 2 CDEs and linked datasets. They will be limited datasets, which include dates and Zip codes (will not include name, street address, contact information, medical record numbers, SSN or any other direct identifiers). Zip codes from projects where participants can be easily identified will not be included in the datasets.

Separate SAPs will be developed for cross-consortium data analysis objectives. Final analytic datasets are dependent on the objectives described in those SAPs. Therefore, additional modifications to these core datasets may be necessary and will be described accordingly.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domain** | **Description** | **CDE Type**  | **Data Type** |
| Sociodemographic | Participant socio-demographic information | Tier 1  | Participant-level |
| Covid test | Information related to COVID-19 tests collected by study teams such as test types and results  | Tier 1 | Record-level |
| Testing | Information related to testing such as accessibility and impact  | Tier 1  | Record-level |
| Symptoms | Participant symptoms as part of testing procedure  | Tier 1 | Record-level |
| Medical history | Participant medical history  | Tier 1  | Participant-level |
| Alcohol and tobacco | Participant history of alcohol and tobacco use  | Tier 1  | Participant-level |
| Location | Participant location  | Tier 1 | Participant-level |
| Vaccine acceptance | Participant view of vaccine acceptance  | Tier 1  | Participant-level |
| Housing/Employment/Insurance | Information related to participant’s housing, employment and insurance status  | Tier 1 | Record-level |
| Work PPE and distancing  | Participant access to work PPE and ability to practice distancing  | Tier 1 | Participant-level |
| Health status | Participant health status  | Tier 1 | Participant-level |

# Core Analytic Dataset Data Dictionary

A link to the Core Analytic Dataset data dictionary will be provided here when available.

# SAP Template

The statistical analysis plan (SAP) template provides a framework for all information needed to conduct the analysis outlined in the protocol. This is a technical document that describes in detail the statistical techniques used by study teams for research analysis. The template can be found in this [public Box folder](https://duke.app.box.com/s/jml32a0z2wcjgyxgy9nlci51mzbeud1t/folder/163435278108).

# Other CDCC Resources

1. [RADx-UP Common Data Elements](https://radx-up.org/research/cdes/)
2. [RADx-UP Data Tool Kit](https://myhome.radx-up.org/cdcc-resources/data-toolkit/)
3. [CDCC Data Submission Guidance](https://myhome.radx-up.org/RADx-UPDataSubmissionGuidance25MAR2022.pdf)
4. [States and Territories Served by RADx-UP Projects](https://radx-up.org/research/projects/interactive-map/)
5. [RADx-UP Data Dashboard Link](https://myhome.radx-up.org/datadashboard/)
6. [RADx-UP Area-Level Data Linkage Dashboard Link](https://myhome.radx-up.org/area-data-linkage/)
7. [Weekly Updates Dashboard](https://myhome.radx-up.org/insights/)