# PUTTING HEALTH CARE DATA ON THE MAP

*Advancing Health Care Equity Through Enhanced Geographic Context in Health Care Administrative Datasets*

MAY 2024

Published by HUMAN SERVICES RESEARCH INSTITUTE
& CENTER FOR IMPROVING VALUE IN HEALTH CARE

## ABOUT HUMAN SERVICES RESEARCH INSTITUTE

Human Services Research Institute (HSRI) is a mission-driven nonprofit organization dedicated to supporting federal, state, and local agencies in providing person-centered, evidence-based, and integrated services to individuals, families, and communities. To drive system change in population health, HSRI collaborates with clients and partners to improve the quality and use of health care data by developing and implementing transparent processes that manage robust health data systems, enhance the reliability and usability of the data, and transform the data into actionable information that researchers, policymakers, and others can use to improve health equity. Learn more at **hsri.org**.

## ABOUT CIVHC

The Center for Improving Value in Health Care (CIVHC) is an independent nonprofit that equips partners and communities in Colorado and across the nation with the resources, services and unbiased data needed to improve health and health care. As the designated administrator of Colorado's All Payer Claims Database (CO APCD), CIVHC oversees the collection of health care claims from Colorado's public and private health care insurers and uses that information to promote price transparency, inform policy, advance health equity, conduct research, and much more. CIVHC is objective, solution-oriented, and maintains the highest integrity in its work. Learn more at **civhc.org**.

# Geocoding | Advancing Health Care Equity Through Enhanced Geographic Context in Health Care Administrative Datasets

## Why Do We Need Geocoding?

Access to health care services and effective use of the health care system is necessary for healthy communities and individual health outcomes, and can vary greatly by community and region. There is growing evidence of the relationship between socioeconomic and other social factors and health care outcomes. For example, living in economically disadvantaged neighborhoods has been associated with higher hospital readmission risk and mortality for heart failure.[1] For this reason, understanding disparities in health care related to demographic and social factors has become a priority for public health entities, state agencies, and health care organizations.

In order to conduct analyses in support of addressing disparities in health equity, the Center for Improving Value in Health Care (CIVHC), administrator of the Colorado All Payer Claims Database (CO APCD), in partnership with Human Services Research Institute (HSRI), CIVHC's APCD data manager who also provides analytic, reporting and data enhancement support, began the process to geocode addresses in the CO APCD in 2020. Geocoding, the process of assigning specific latitude and longitudinal coordinates to addresses, enables CIVHC and HSRI to conduct community-level analyses and bring in external sources of data to understand the correlation between social factors and health access, and more. This paper describes lessons learned in the process of geocoding the CO APCD including the benefits of geocoding, privacy considerations, tool selection, and processing.

[1] Shirey, T. E., Hu, Y., Ko, Y. A., Nayak, A., Udeshi, E., Patel, S., & Morris, A. A. (2021). Relation of Neighborhood Disadvantage to Heart Failure Symptoms and Hospitalizations. *The American journal of cardiology, 140*, 83–90. https://doi.org/10.1016/j.amjcard.2020.10.057. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8764641/

# Background

There are considerable challenges to identifying health inequities, implementing new systems, and monitoring progress toward achieving health equity. Administrative health care claims data—records of health care service and prescription drug payments collected by commercial and public health insurance payers— has very limited ability to identify individual-level information on social characteristics (like education, income, or housing characteristics). Many payers simply do not collect this information in their systems, and incorporating these data elements from other sources takes time and is resource intensive. Area-based measurement—describing larger population groups and geographic areas, such as percentage of population that is unemployed in a given neighborhood—is a promising alternative to help understand the socioeconomic environment surrounding health care cost, utilization, and quality.

Area-based social measures are powerful because they make effective use of patient address data while maintaining privacy and can be linked with patient-level data. Patient addresses are routinely collected and updated for administrative purposes, and these addresses can be geocoded, which assigns specific latitude and longitude coordinates to each address. Once addresses are geocoded, area-based information from other public or non-public data sources can be linked with the health care administrative data at different levels of geographic aggregation—such as state, county, city, and census tract—to characterize the social and environmental aspects of the geographic regions.

Area-level data in this case is used to describe areas with particular characteristics that might be associated with health care information. They are not a substitute for individual-level attributes, such as having firsthand information on a patient's education level, housing situation, employment status, or income level. However, models suggest that for at least some health outcomes, area- and individual-level socioeconomic factors independently and jointly shape the population distribution of disease and health behaviors.

In addition to gaining an understanding of the socioeconomic environment, another advantage of using geocoded data is that it contributes to more precise measurements of distance and travel time between patient residence and where they access the health care system (e.g., hospitals, urgent care, primary care offices, specialty practices, pharmacies, ambulance providers) to better understand the accessibility of services.

As an alternative to precise geographic coordinates, analysts are able to calculate distances between the center of ZIP codes, which is largely populated for patients and providers in administrative datasets and more readily available for all ZIP codes in the nation. However, using the center of ZIP codes only provides an approximation of location and distance, as opposed to the true distance between residence locations and health care service providers. The approach that relies on ZIP code centroids is not suitable for the evaluation of availability and access using travel time- or distance-based measures that require more precise locations. For example, evaluating which areas are not reachable by ambulance within a 25-minute

drive, or the share of rural residents that can access pediatric behavioral health providers within 60 miles of their home.

In these and other similar scenarios, precise geographic coordinates obtained through geocoding in conjunction with advanced geographic analyses tools that model distance and travel time with existing road systems are necessary to conduct robust provider network adequacy analyses.

# Considerations Before You Get Started

Considering the benefits of geocoding that have been highlighted above, it is important for health care data administrators to consider a few other crucial points before embarking on developing a geocoded data infrastructure.

## Privacy Concerns

Privacy should be the No. 1 consideration for selecting a geocoding tool or service. Consider asking for information about the degree to which the geocoding tool supports HIPAA compliance, which is typically described in the technical specifications for data hosting, user access controls, authentication, etc.

Even when using a HIPAA-compliant geocoding tool, precautions should be taken throughout all data transfer steps to ensure there is minimal risk involved in the process. Addresses extracted from the secure data warehouse environment should be stripped of any identifiers that could directly link back to a person or claim in the health care records (also known as de-identification). The content of the input file should be kept to the minimum necessary for processing (i.e., the file contains only the address elements).

When working with information relevant to patients, or insured individuals, as opposed to providers, detailed street address elements are considered Personally Identifiable Information (PII), or direct identifiers. These elements are not releasable or accessible to most data users. Even for data users who are approved to access these data elements, data handling must follow highly secure procedures.

Therefore, receiving data with geocoded address information is often not feasible for all data users. The geography elements that are typically available for users of de-identified data sets are either the first three digits for ZIP codes, or county-level information. Both of these options represent low granularity geographic areas that would obscure important variation in resident characteristics and their health care outcomes.

Storing and accessing geocoded output or derived data elements results in additional privacy concerns. Geographic coordinates hold as much detail or identification risk for a person as a street address, depending on precision and accuracy of the geocoded output (concepts described further below). The presence of geocoded data elements dictate the database schema where the output will be loaded (i.e., a separate database schema with PHI data and special access restrictions).

We recommend ensuring that privacy implications for both input and output information are fully understood and acknowledged by all internal and external data users, and are considered when making decisions about the storage and release of data, and throughout the geocoding process.

## Selecting the Optimal Geocoding Solution

We used a geocoding tool to process both member/patient addresses and provider addresses. Although the latter were primarily sourced from a publicly available dataset and not subject to the same privacy concerns, the same steps were followed in the geocoding process for both sources.

The primary objective for our initial geocoding use case, bringing in social factor data at the census tract level, was to identify a *forward* geocoding tool (i.e., looking to transform address details into geographic coordinates, rather than to transform geographic coordinates into physical addresses). This approach starts with detailed address information for patients/members and health care providers to obtain geographic latitude and longitude coordinates. Several options are available for geocoding address information and it can be a daunting task for health care database or APCD data administrators.

While data administrators may be inclined to search for a tool that provides a one-time geocoding process, finding a trustworthy tool for ongoing, periodic geocoding updates offers continuity for analyses and other use cases.

The most important considerations for tool selection based on our perspective of working with CO APCD addresses are outlined below.

### Capacity To Process High Volume of Addresses

One of the most important factors in the selection of a robust geocoding tool for a large health care data set is to identify a tool that can process millions of addresses in a streamlined way. For our purposes, we needed a tool capable of loading millions of records at the same time.

### Integration of Value-Add Data Elements

Consider selecting a tool that can provide further value-add information associated with the addresses, such as the assignment of standardized geographic identifiers from the U.S. Census Bureau, like the census tract Federal Information Processing Series (FIPS) code. These identifiers can then be used to link to area-based population statistics, such as the percent of residents in a census tract that are below poverty line.

### Processing Cost

The cost of using a tool on a regular basis (annually, at a minimum) may be prohibitive, depending on the tool selected. It is important to define how much data will need to be processed and how often it will need updating before selecting a tool. Separately, it is important to assess internal time and staff resources needed to coordinate the geocoding process, monitor progress, load, download, transfer the data, and maneuver the data through the required steps. For our purposes, we decided on an annual update frequency, and determined we could process all of the claims within a month timeframe.

Given our requirements, we selected a tool that allowed for unlimited processing of addresses for a single monthly fee. With the tool we used, some of the input files took hours to process, some took days, and some files had to be reprocessed due to a variety of reasons. In spite of some of the unexpected glitches, having access to the tool for one month was sufficient and provided sufficient room for error. We found that to process the complete set of addresses described in the Geocoding Input section below, a minimum of 12 consecutive calendar days (two business weeks) was necessary.

### Limitations to Output Uses

It is important to understand whether the company licensing the tool you select permits the use of geocoding output for all of your anticipated uses. For example, you may need to determine if the output can be used by and leased to other parties, and if so, whether there are any specific conditions to consider for certain types of users. You will also want to understand if there are any additional fees or costs associated with other use cases.

### Tool Testing and Documentation

For our selection process, we started by reviewing summaries of tools other researchers or data users have used.[2] One tool, Geocodio, was listed as one of the best options in some of the online geocoding tool reviews, and after researching further, we confirmed the tool met our criteria for processing capabilities and price point.

Geocodio allowed us to perform multiple processing tests with public address files using their standard product for free, subject to their daily lookup limit. We tested a variety of scenarios for input address data quality, such as incomplete address information, misspelled location names, and extraneous address information (e.g., name of business place before the street address). Based on our preliminary tests, Geocodio's performance met our expectations and returned the expected geographic coordinates and additional output elements even for less straightforward addresses. For example, "123 Main," "123 Main St," "123 Main Street," and "123 Maine St" are addresses with text variations that represent the same location to a human interpreter, where the same pair of geographic coordinates was returned. Through testing we confirmed that the tool's processing logic was capable of successfully finding matches and that addresses were matched to the same location as expected. We then spot-checked using a different online tool to confirm that the returned coordinates corresponded to the original address. Based on the positive results, we adopted Geocodio to process the complete set of addresses from the data warehouse.

Since tool testing is important, we recommend creating a set of addresses that cover as many completeness and quality scenarios as possible, and performing an assessment to confirm that the output matches expectations. Some questions to guide the assessment could include:

1. Do input addresses that represent the same location yield the same geographic coordinates in the output?
   - If not, how far apart are the locations they represent?
   - Did they receive a different geographic identifier or the same identifier (e.g., Census Bureau census block, census tract identifiers)?
2. Which data quality issues (e.g., street spelling, abbreviations) resulted in erroneous locations or output coordinates with low accuracy?
3. Are there particular input address details that are effectively ignored by the geocoding process (e.g., suite or apartment numbers, last four digits of nine-digit ZIP codes)?

Throughout the testing process we recommend documenting:

- The tool-testing results,
- Decisions made to select a particular geocoding tool,

[2] The Public Health Disparities Geocoding Project. *Geocoding*. Retrieved from https://www.hsph.harvard.edu/thegeocodingproject/geocoding/

- Criteria to determine acceptable levels of geocoding precision and accuracy, and
- Versioning and potential changes in the geocoding algorithm (if there is transparency from the source) between rounds of processing.

Another important item to consider is the geographic coordinate system used by the tool. This is relevant for any future mapping work using the geocoded addresses. For example, Geocodio uses the standard World Geodetic System 1984 (WGS84) geographic coordinate system.

Investing time to document the internal process and resources involved in performing address geocoding on a regular basis will support future iterations or comparing available geocoding tool options.

## The Geocoding Process

As of spring 2024, our teams have successfully completed three rounds of geocoding of CO APCD addresses. We have identified and completed some small refinements to our original processes and subsequent implementation of the geocoding output in the CO APCD.

Exhibit 1: Geocoding Process Steps shows the typical geocoding process when using a browser-based geocoding tool external to the data warehouse environment.

The process may be different depending on your data warehouse setup, or if you decide to use a different type of tool. However, the observations and insights in this document aim to be independent of the type of tool selected.

**Exhibit 1. Geocoding Process Steps**

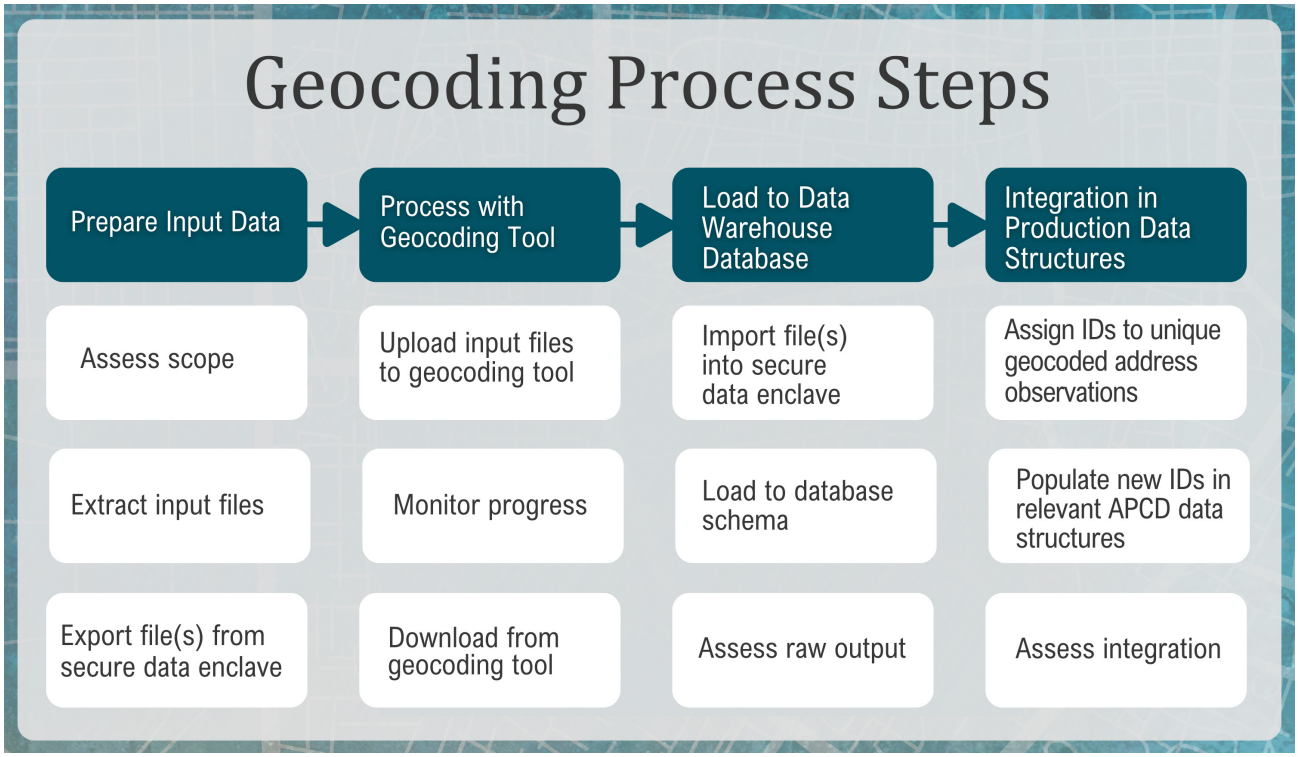
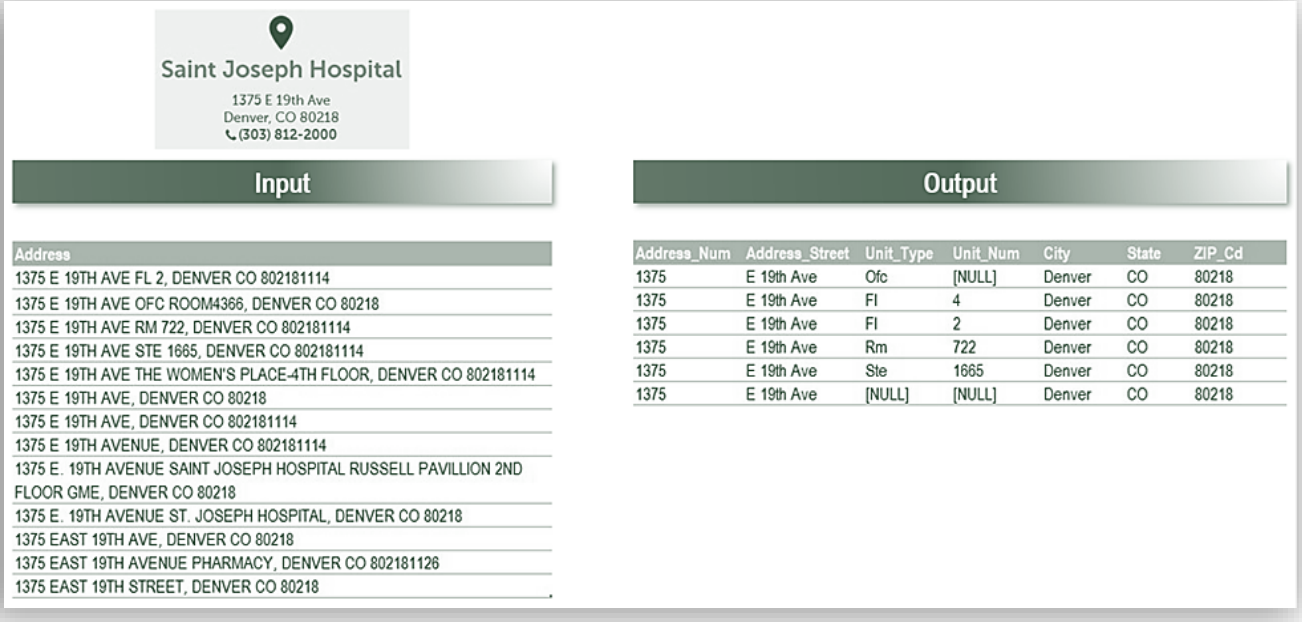


**Exhibit 2. Examples of Input Address Observations and Output Standardized Addresses**




**Input**

| Address |
|---|
| 1375 E 19TH AVE FL 2, DENVER CO 802181114 |
| 1375 E 19TH AVE OFC ROOM4366, DENVER CO 80218 |
| 1375 E 19TH AVE RM 722, DENVER CO 802181114 |
| 1375 E 19TH AVE STE 1665, DENVER CO 802181114 |
| 1375 E 19TH AVE THE WOMEN'S PLACE-4TH FLOOR, DENVER CO 802181114 |
| 1375 E 19TH AVE, DENVER CO 80218 |
| 1375 E 19TH AVE, DENVER CO 802181114 |
| 1375 E 19TH AVENUE, DENVER CO 802181114 |
| 1375 E. 19TH AVENUE SAINT JOSEPH HOSPITAL RUSSELL PAVILLION 2ND FLOOR GME, DENVER CO 80218 |
| 1375 E. 19TH AVENUE ST. JOSEPH HOSPITAL, DENVER CO 80218 |
| 1375 EAST 19TH AVE, DENVER CO 80218 |
| 1375 EAST 19TH AVENUE PHARMACY, DENVER CO 802181126 |
| 1375 EAST 19TH STREET, DENVER CO 80218 |

**Output**

| Address_Num | Address_Street | Unit_Type | Unit_Num | City | State | ZIP_Cd |
|---|---|---|---|---|---|---|
| 1375 | E 19th Ave | Ofc | [NULL] | Denver | CO | 80218 |
| 1375 | E 19th Ave | Fl | 4 | Denver | CO | 80218 |
| 1375 | E 19th Ave | Fl | 2 | Denver | CO | 80218 |
| 1375 | E 19th Ave | Rm | 722 | Denver | CO | 80218 |
| 1375 | E 19th Ave | Ste | 1665 | Denver | CO | 80218 |
| 1375 | E 19th Ave | [NULL] | [NULL] | Denver | CO | 80218 |

## Geocoding Input

Our objective was to geocode address information for both people as well as providers in the CO APCD. The latter involves all primary and secondary addresses available in the National Plan and Provider Enumeration System (NPPES), totaling roughly 5 million distinct address observations. There are approximately 15 million address observations for distinct people in the CO APCD that get processed on an annual basis, which include addresses from 2013 records through the most recently submitted CO APCD records.

Generally, one can expect that geocoding tools would, at a minimum, handle variations such as "street" and "st" and variable punctuation or capitalization. The best tools should also be able to handle misspellings and ZIP codes that are missing a digit, but it may be dependent on the completeness and validity of the address as a whole.

As an example of the tool processing ability, Exhibit 2 displays a set of addresses present in NPPES that reference the location of the Saint Joseph Hospital in Denver, all of which received the same geographic coordinates through geocoding. In addition to coordinates, the geocoding tool also parsed these input addresses into distinct address elements, resulting in a smaller set of standardized addresses in the output than initially included in the input data.

Given the volume of records and the fact that the tool tests we performed with the uncleaned addresses produced good outcomes, our teams were comfortable applying only minimal cleaning to input addresses (prior to deduplicating the list) in order to remove some variation and streamline the list of distinct address observations to be processed. These steps involved:

1. Removing particular non-alphanumeric characters (e.g., return, new line characters),
2. Swapping order of address elements as needed, and
3. Removing leading or trailing space characters.

Prior to conducting the geocoding process, there are a few input address scenarios described below to be aware of that require an assessment of volume.

### Addresses With Missing Street Details

One can expect that some addresses will have missing information. For example, some may include only the city and state, or only city, state, and ZIP code, with no street details available. In the CO APCD data, we found that 100% of the Medicare beneficiary data from CMS had missing address information, which may be the case for other states who receive similar datasets.

In such a scenario, regardless of the geocoding tool used, the output coordinates cannot have more granular information than the input. In fact, the output will most likely contain the geographic coordinates of the centroid—or geographic center—of the city or of the ZIP code tabulation area (ZCTA). In such instances, some geocoding tools may provide better results if the nine-digit ZIP code is supplied in the input address, as opposed to the five-digit ZIP code.[3] Some tools, such as the Geocodio tool, do not have this capability built in their system, and any nine-digit ZIP code is processed as a five-digit ZIP code.[4]

### Post Office Box Addresses

There are special considerations that apply to addresses that represent Post Office Boxes (PO Boxes). These addresses refer to a box at a post office where a person or business has mail delivered. PO Box addresses typically only contain the PO Box number, city, and ZIP code, with the city and ZIP code representing the location of the post office.

Occasionally, PO Box addresses include the detailed street address; however, if a PO Box is present, it typically reflects the post office location, rather than a residential address. Data users should monitor the percentage of records with PO Box addresses in their analyses and consider including a cautionary note informing data users about the volume of people in the database that may have an inaccurate geolocation. In the case of the CO APCD, we found that approximately 6% of the input address observations represent a PO Box as opposed to a residential address.

While PO Box addresses exist in both rural and urban counties, in the CO APCD, a higher proportion are in rural counties. Depending on the type of analysis, data users may want to consider removing records with PO Box addresses from analyses. To help support this exploration, you may consider constructing a binary flag with relatively straightforward logic that is applied to the input street detail information and considers all possible spellings of PO Box (e.g., PO Box, P.O. Box, Post Office Box, Postal Box), including misspellings. You may also consider distributing this flag—or summary information about volume of PO Box addresses—to data users, since data users are likely not going to have access to the detailed address information to perform this assessment on their own.

It is also important to note that, for the most accurate geographic analyses, patient addresses should represent their home address. The extent to which patient addresses represent something other than their home address may not be fully detectable; however, it is safe to classify all PO Box addresses in the mailing address

[3] This is not the case for Medicare Original beneficiary data from CMS. See https://resdac.org/cms-data/variables/zip-code-beneficiary

[4] Geocodio does offer USPS ZIP+4 codes along with some deliverability indicators for US addresses, as part of the output data elements (https://www.geocod.io/guides/zip4/), which can support the development of additional processing logic for 9-digit ZIP codes in the data warehouse, post geocoding.

category. With further exploration using the geocoding output, it may be possible to detect other mailing addresses by looking at the characteristics of the census tract to which a person address is assigned.

## Geocoding Output and Value Adds

The information returned after processing addresses with a geocoding tool typically falls into three categories, as shown in Exhibit 3.

In our case, the tool allowed us to request the inclusion of geographic identifiers—the highest granularity being the 15-digit FIPS code—at the census block level; state (two digits) + county (three digits) + census tract (six digits) + census block (four digits); metro and micropolitan area name and identifiers, combined statistical area name and identifier, state legislative district numbers, and congressional district numbers. More options of value-add elements were available if needed, such as population statistics linked to the respective census geographies.

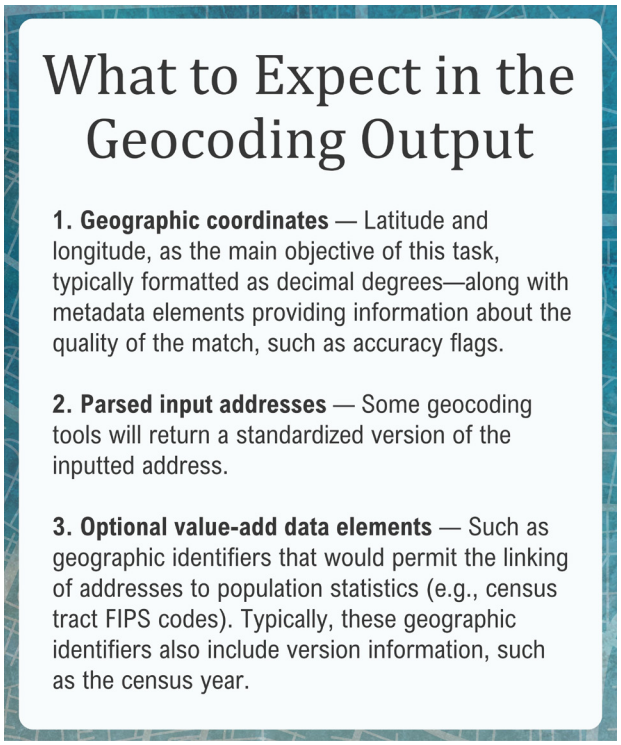
Data users should also be paying close attention to the precision of the returned geographic coordinates and to the metadata elements that provide information about the quality of the match—and consequently about any of the associated output columns.

### Geographic Boundary Version

Part of the geocoding process planning steps should include an assessment of what area boundary versions are available for the value-add geographic identifiers that can be acquired. It is important to proactively identify data linkages and data releases that need geographic identifiers, and ask questions such as:

- Will the geocoded dataset be linked to the most recently released statistics or to historical statistics—or both?
- What years of data would therefore be needed?

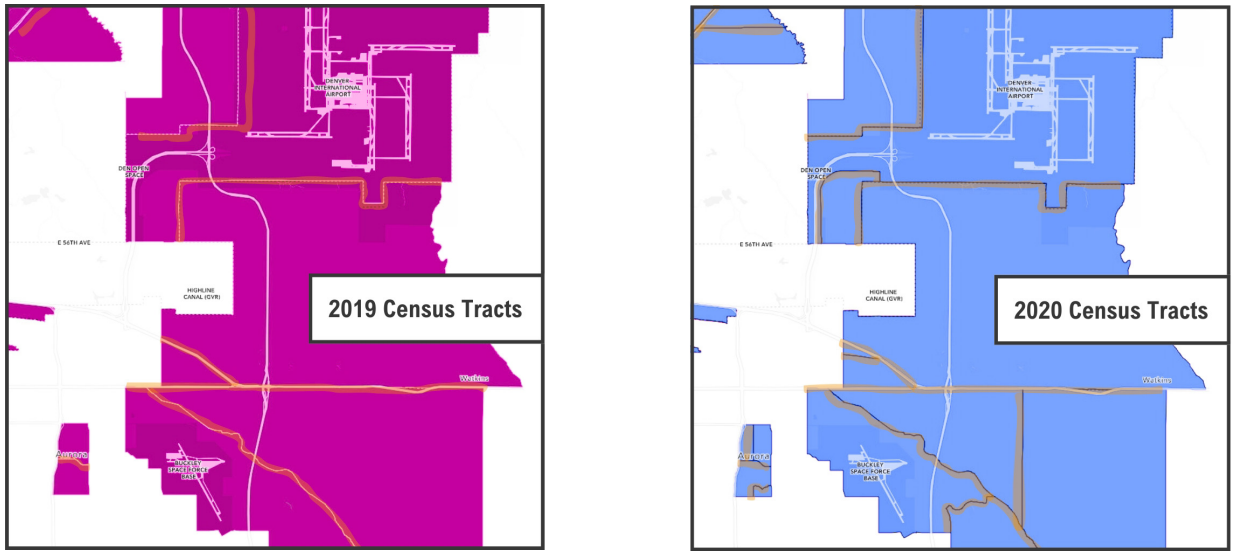
**Exhibit 3. What to Expect in the Geocoding Output**




The tool we used allows users to select from multiple geography identifier versions. For example, for census identifiers including census tracts, users can pick identifiers "as of the 2019 census year" or "as of the 2020 census year," among a longer list of versions.

The boundaries of geographic areas such as census tracts (and their assigned unique identifier, FIPS codes) are expected to remain fairly stable across versions released between decennial census years, such as for 2010 through 2019 Census Bureau data. However, substantial changes to the area definition can be expected once every decade since the Census Bureau reevaluates and makes changes to boundaries in response to population changes. For example, census tracts with population increases may be split into two or more tracts, those with population decreases may be combined, and so on. Such changes of areal definitions mean changes in the list of assigned FIPS codes (i.e., new spatial units receive new identifiers).

**Exhibit 4. Comparison of 2019 to 2020 Census Tracts (example from the Denver area)**




Note: Exhibit 4 shows examples of the Denver-area census tracts that were split into two or more tracts. Note the 2019 tract boundaries highlighted in gold on the map at left; then, compare 2020 tract boundaries highlighted in gold on the map at right. In the 2020 census tract map (at right), observe additional tract boundaries highlighted in the areas surrounding Denver International Airport and Buckley Space Force Base, as well as within the city of Aurora. Source: https://www.esri.com/arcgis-blog/products/arcgis-living-atlas/mapping/acs-2016-2020-updated-boundaries/

This is important to know because the linkage between geocoded data and external data sources can produce mismatches if separate data sources use FIPS code versions that are before and after such changes are implemented. For example, data releases based on the census year 2020 definition have a relatively large share of census tract FIPS codes that do not appear in data releases based on the census year 2019 definition, whereas the latter will likely have the same or almost entirely the same set of census tract FIPS codes as data releases based on the census year 2018 definition. Exhibit 4 displays an example from the Denver region that shows a set of census tracts that had boundary and FIPS codes changes between the 2019 and 2020 versions. Census tracts not displayed in color have the same FIPS codes in both versions.

## Precision

Precision of the geographic coordinates refers to the number of digits that are available in the geocoded output, for coordinates represented as decimal degrees (Exhibit 5).

**Exhibit 5. Geographic Coordinates Formatted as Decimal Degrees: Precision and Scale**

| Decimal places | Decimal degrees | Object that can be *unambiguously* recognized at this scale |
|---|---|---|
| 0 | 1.0 | Country or large region |
| 1 | 0.1 | Large city or district |
| 2 | 0.01 | Town or village |
| 3 | 0.001 | Neighborhood, street |
| 4 | 0.0001 | Individual street, large buildings |
| 5 | 0.00001 | Individual trees, houses |
| 6 | 0.000001 | Individual humans |
| 7 | 0.0000001 | Practical limit of commercial surveying |
| 8 | 0.00000001 | Specialized surveying (e.g., tectonic plate mapping) |

Source: https://en.wikipedia.org/wiki/Decimal_degrees

The Geocodio output provided six decimal places for coordinates. This level of precision is more than sufficient for most health care data use cases. Other geocoding tools may return seven to 10 or more decimals. Whether one picks the seven or 10 decimals for geographic coordinates, distance estimates between two points or analyses based on geographic areas such as census tracts or blocks would have the similar results.

### Accuracy

Geocoding processing tools should also provide metadata elements describing how input addresses match up against the databases being used by the tool. Metadata may contain biases since it is created by the tool owner, so it is best practice to run a sample by one or several other geocoding tools to compare not only the geographic coordinates returned, but also match quality indicators. The tool we use includes two indicators of accuracy[5] in the output, Accuracy Score (scale: 0 to 1) and Accuracy Type—rooftop, point, county, place, state, intersection, etc.

### Custom Output Quality Indicators

To further support end users of the geocoded data, you may consider creating a value-add indicator identifying the most reliable geocoded output. After thoroughly reviewing the output, geocoding tool documentation, and guidance to users, and also comparing some test address output against a different tool often used in the industry, we created a binary flag indicating whether the address has been linked to a high-quality geocoding output based on whether the respective address meets all of the following conditions:

(a) Input and output address information includes a ZIP code and the input ZIP code is the same as the output ZIP code,
(b) The Geocodio output Accuracy Type is either "rooftop" (i.e., on the exact parcel) or "range interpolation" (i.e., generally, in front of the parcel on the street), and
(c) The Geocodio output Accuracy Score is above 0.8.

Some data users may prefer to rely directly on the accuracy indicators provided in the output, so it may be optimal to provide these data elements as well.

## Implementation in the APCD Data Warehouse

The details described here are consistent with our data warehouse implementation; however, they may vary based on the specific APCD or other health care dataset that geocoding will be applied to.

Once the geocoding processing is complete, the output can be imported into the data warehouse, and decisions need to be made regarding linkage with APCD data. For example, to associate claims or eligibility records with a particular census tract ID, a decision needs to be made between adding the respective data element to the production data structure, or adding a record identifier that links the APCD record back to the data structure that contains the census tract ID. As mentioned before, privacy concerns would inform the data structure setup. In either case, best practice would be to create Geocoded Address IDs specific to the output data structure.

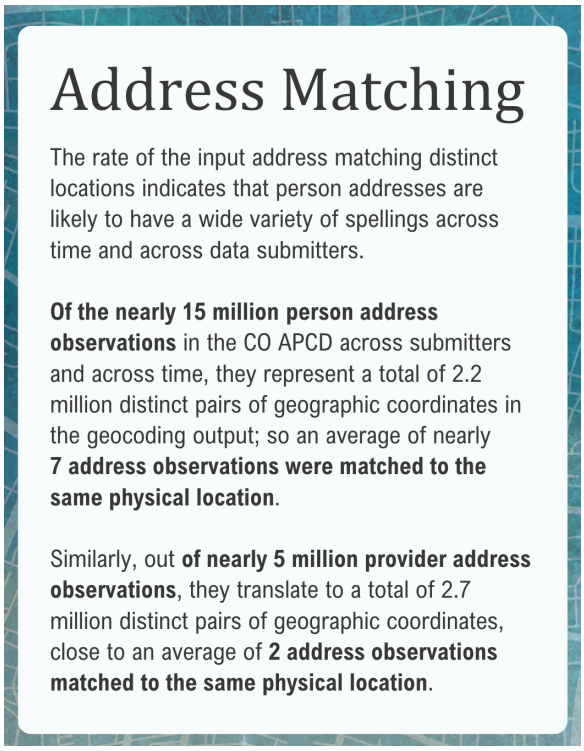
Another consideration is that the periodical processing of addresses requires the implementation of a versioning system for the geocoded output and ensuring that the APCD data will link to the most up-to-date version available for the respective records.

If you are implementing an annual geocoding

[5] More information about these data elements is available here: https://www.geocod.io/guides/accuracy-types-scores/

processing, with each APCD data submission after the annual processing, there will be a higher volume of claim and eligibility records that do not have a geocode match. This is expected because submissions will include new addresses. A portion of the new addresses may represent locations that are already present in the geocoding output, and it is possible to design a custom interim process to look for additional matches. However, the respective process would likely involve manual exploration and pairing, and, depending on the volume of new addresses, it may not be feasible to incorporate this step into your regular data release cycle. In this case, it may be best to wait until the next annual geocoding iteration, when all addresses—historical and new—in the data warehouse would be included in the input file.

**Exhibit 6. Address Matching**




# Recommendations

## For APCD or Other Health Care Data Administrators

An investment in applying geocoding to person and provider addresses has many benefits for data linkages and more precise geographic analyses, including supporting health equity analyses. However, using the geocoded information comes with some commitments, responsibilities, and caveats that must be accounted for. Following are some benefits and suggestions based on our experience with CO APCD data and the geocoding process.

1. **Improving Submissions.** Going through the geocoding process can provide a great deal of information to **support the overall goal of improving address information submitted to the APCD**. Geocoding requires taking a closer look at submitter-level patterns of lower accuracy addresses, historical and current, which reveal opportunities to connect with submitters and improve the information in future submissions.
2. **Address Matching.** One advantage of geocoded data is that it can help match addresses in the data warehouse that represent the same physical location. Consider the following scenario: Person A has coverage from more than one payer or more than one plan for a period of time and has two or three different addresses (with either just a slight variation in spelling or a completely different address) in their eligibility records for December of a given year. Geographic coordinates obtained for these addresses can support an exploration of whether the December addresses are all tied to the same physical location or if they represent a different location, through a calculation of distance between the respective pairs of coordinates. Alternatives exist but are

suboptimal; mathematical "distance" between text strings, or manual exploration of address values, are some examples.

3. **Address Standardization.** Depending on the tool used, the standard address elements returned in the geocoding output may support additional analyses on the quality of the input address information, although we would advise caution—the address-parsing algorithm used by the tool may not universally yield the same standard cleaned address elements for address observations that are otherwise geocoded to the same location.

4. **User Guidance.** Consider creating succinct and clear guidance for all future users of data elements resulting from the geocoding process that summarize the decision points described above and are pertinent to the data source or even tailored to the specific use case of the data requestor. After that, ensuring that guidance documents are kept up to date and in line with the current address processing cycle is a good practice, as is distributing guidance documents alongside data files.

## For Analysts Working with Geocoded Data

There are a couple of important takeaways for analysts as well. First, it is important to remember that geocoded person data is PII and must be handled appropriately. For example, if an analyst were to make maps using geocoded data, they must never display the actual addresses in the map.

Second, limitations inherent in the input address data will also be present in the output, whether related to missing street detail for certain historical years or for certain data submitters, or clear evidence that addresses are mailing addresses rather than residential addresses, when they include PO Box details. Whether your goal is to link your data to the Census Bureau data or other external datasets, or to perform travel distance calculations, it is important to keep in mind that the availability of a census identifier or another location identifier on your eligibility or claim records must be used in conjunction with the geocoding accuracy flags available in the output, or any other custom value-add flags created. How you are using the addresses in your analysis will inform your decision to include or exclude lower quality addresses or account for their presence in some other way.

We recommend thinking through these aspects on a case-by-case basis and consider assessing, summarizing, and including some information about geocoding accuracy and PO Box volume in the methodology notes of your report.

If the most current person address has a low level of accuracy and high accuracy is of utmost importance for the analytic use case, analysts should strongly consider developing a data structure that contains person-level address assignments for a specific period of time such as at the year level, through the use of geocoded output. For example, it is our experience that Medicare Original beneficiary addresses lack street-level detail; however, a substantial portion of these members may have more detailed addresses with higher accuracy for the same time frame either from Medicare Advantage records that are submitted through commercial payers, or, if they are dual-eligible individuals, through their Medicaid eligibility records.

Finally, when using mapping software for drive time or distance analyses, it is important to use the same geographic coordinate system in future work as the one used to geocode the address data. As mentioned, Geocodio uses WGS84.

# A Summary of Considerations in a Geocoding Project

- Ensuring HIPAA compliance: geocoding tool, addresses processing, data release, and data reporting
- Decisions to use and to release the geocoded data to third parties must be based on specific data release policies applicable to your dataset
- Tool selection depends on affordability, capacity, and output results
- Analysts should consider using indicators of geocoding accuracy to support analytic decisions
- Integration of value-add data elements (e.g., census geography identifiers such as census tract FIPS codes) to support data linkage with area-based population statistics; special attention needed to select the optimal version of geographic identifiers
- Using geocoded data to provide contextual information for health care performance analysis, but also to perform quality checks on data received by submitters (e.g., by using standardized addresses, matching addresses to unique physical locations)

HSRI



CIVHC
CENTER FOR IMPROVING VALUE IN HEALTH CARE