TRACKING PROGRAM GEOCODING STANDARDS: BACKGROUND AND RATIONALE

January 2023

Tracking Program's Sub-County Efforts

Public health, including environmental health, is a local issue, and local data are needed to properly identify and address these issues. The Centers for Disease Control and Prevention's (CDC) Environmental Public Health Tracking Program (Tracking Program) is moving toward building a system of sub-county data to enhance the spatial resolution of data currently available in the National Environmental Public Health Tracking Network (Tracking Network). Currently, most health data displayed on the Tracking Network are state- or county-level data. This project has four components:

1) Many health outcomes cannot be displayed at the census tract level for reasons of confidentiality and stability due to too few cases or people in an area. **Geographic aggregation** will be used to combine census tracts into standardized sub-county geographies to allow for data to be displayed at a finer geographic scale than county-level data.

2) Even with aggregation, suppression will be needed to protect confidentiality. The Tracking Program's current **suppression rules** cover county-level data. Therefore, the Tracking Program will evaluate suppression rules used by grantees and other data stewards for protecting sub-county data to develop new suppression rules specifically for sub-county data.

3) Evaluations will be conducted to better understand how the use of different datasets for **population estimates** (e.g., American Communities Survey 5-year estimates and decennial census) affect rates. Census tract population data and expected census tract health outcome counts will be used to develop a framework for evaluating different datasets of population estimates.

4) Census tracts will be used as the Tracking Network's foundational sub-county geography. As such, the Tracking Program will collaborate with recipients to develop **geocoding standards** for transforming address-level health data to census tracts.

All of the efforts from this project will expand the capability and utility of the Tracking Network, meeting the increasing demand for finer geographic resolution data. Enhancing the Tracking Network in this manner requires collaboration across many parties to advance environmental public health surveillance. This document focuses on the fourth component of the Tracking Program's sub-county efforts and describes how the Tracking Program's Geocoding Standards (Geocoding Standards) were developed. The Geocoding Standards document describes how to geocode addresses to a census tract in a standardized manner for the Tracking Program or for any program that wishes to present census tract-level data or any sub-county-level data. The following sections are a brief introduction to geocoding (Background), a review of the Tracking Program's sub-county pilots involving geocoding, and a list of decisions that were made during the writing of the Geocoding Standards document.

Background

Automated geocoding is the "practice of using a geographic information system (GIS) to match an address to a street name and address range in a digitized street reference map."[1] This can be done at multiple levels, including matching an address to a state, a county, a town, a census tract, or even a point (longitude and latitude). Researchers who are interested in investigating environmental exposures or health outcomes over space may geocode addresses from health records or other sources to a point or region on a map. The Tracking Program currently accepts case counts that are aggregated to the county or state level. Previously, the Tracking Program did not request information on how recipients geocoded their data. Because of the transition from collecting county-level case counts to sub-county-level case counts, the Tracking Program reviewed existing information about geocoding processes and created a set of best practice guidelines for the Tracking Program.

The Tracking Program identified three commonly used GIS software tools: ArcGIS, Centrus/MapMarker, and the Texas A&M Geocoder. These programs assign an address to a point or polygon on a map based on matching an address to an element in an underlying database. If an input address is not an exact match to an address in an underlying database, the software may place a point by an informed guess based on address string components (interpolation) or increase spatial granularity by moving the point to a larger unit like a street segment or polygon centroid.[2] GIS software typically produce output codes for each address, indicating the spatial resolution of the match or match completeness.

Although the three software programs frequently used by recipients might display similar outputs, differences exist between the programs, making direct comparison of results challenging. First, the choice of database and the completeness of the database will affect geocoding results.[2, 3] For example, a database might not include new construction.[4] Some databases contain information on the latitude and longitude of an address to a front door while others may feature rooftop coordinates. Additionally, software users can adjust the sensitivity of address matching and spatial resolution.[5-7] Users can specify settings about how closely an input address needs to match an address in a database before the match is moved to a larger geography (i.e., to a coarser resolution). Software have different processes for handling an input address that does not exactly match an address in the underlying databases. For example, some software will use interpolation to match where an address might be based on other known pieces of the address like the house number or street suffix. Based on programmed settings, a software might "zoom out" until the address can be matched with a centroid if a point match or street segment is not available.

The most significant difference that makes direct comparisons difficult is that geocoding software do not produce identical output metrics. Some outputs include fields for match score, feature match, quality type, latitude, longitude, and census Federal Information Processing System (FIPS) codes, among others. Centrus/MapMarker provides alphanumeric codes to label the output (e.g., ZT5V). ArcGIS and the Texas A&M Geocoder both provide a match score. Match score is specific to each program and is a measure of how well an input address matches an address contained within an underlying database; **it is not a measure of precision or positional accuracy**.[2, 8] A match score of 90 in one program is not equivalent to a match score of 90 in another program. Centrus/MapMarker does not generate a match score; similar fields in this software would be result code or location code.

The Tracking Program is interested in geocoding to the census tract. A review of available literature identified a gap in published information on geocoding to a census tract, particularly as it relates to environmental public health. According to Zhan et al.,[4] the validity of epidemiologic research depends on the positional accuracy of geocodes and the match rate of successfully geocoded addresses. Most of the published geocoding literature

deals with assessing the accuracy and precision of geocoding to point (e.g., roof top, front door, or latitude/longitude) references. Some authors also investigated how match rates, which are the percentage of addresses that are geocoded successfully, might vary based on address characteristics (e.g., rural versus urban areas) or database characteristics. Studies indicate that urban addresses geocode successfully more often than do rural addresses.[3, 9, 10] This difference might have greater influence when small numbers are present, as in the case of rare disease rates calculated from rural census tracts. Missing case counts due to geocoding failure will result in underestimating the true rate of disease at a larger magnitude when the counts are small (e.g., census tracts compared with county).[11]

Currently, there is no measure of geocoding precision or automatic process for ensuring that an address is plotted to a map correctly. Although match score looks like a measure of precision, with a score of 0–100, it is only a measure of how well the pieces of an address match the pieces of another address in an underlying database. The location and result codes by Centrus/MapMarker indicate how the software matched the address and interpolated and placed a point, but they are not measures of precision. Currently, the only method of identifying the accuracy of a geocode is to visit the plotted location and use a global positioning system (GPS) device to manually check the location with the address.

Another significant consideration is determining what constitutes a successful geocode. GIS settings for specificity and address matching can be adjusted to control the spatial resolution of matches. Software will sometimes "zoom out" to a coarser resolution (e.g., from a street segment to a ZIP code centroid) when the software is unable to match the higher resolution geography. The Tracking Program identified two metrics to describe the match quality of a geocode: 1) the spatial resolution of the match and 2) the match score. Details about the Tracking Program's definition of a successful match can be found in the Geocoding Decisions section.

It is not always possible to successfully geocode a record to an acceptable spatial resolution (e.g., a residential address or census tract). In such cases, a researcher may choose to discard these non-geocoded records. However, this might introduce the potential for geographic selection bias in which data might be missing non-randomly from particular geographic regions of the study area.[12] To limit the potential for such bias, researchers may choose to use methods to assign a record to a "reasonable" location, a process known as geographic imputation.[13] The two main kinds of geographic imputation methods are deterministic and stochastic. Deterministic geographic imputation relies on a set of rules to assign a record to an appropriate geographic location based on available demographic and spatial information. Stochastic geographic imputation uses these variables to generate a statistical model of the probability of a record being assigned to the correct location.[14] Studies have shown that choosing an optimal imputation method depends largely on the specific application, and greater complexity does not necessarily produce the most accurate results.[14-17]

Other agencies and organizations, such as the North American Association of Central Cancer Registries (NAACCR) and the National Center for Health Statistics (NCHS), have worked to establish their own geocoding best practices documents. In 2008, NAACCR published "A Geocoding Best Practices Guide," in collaboration with Daniel Goldberg from University of Southern California.[2] Swift, et al.[18] prepared a manuscript for the CDC Division of Cancer Prevention and Control that contained a review of geocoding software. NCHS addressed geocoding practices in its United States Small-Area Life Expectancy Estimates Project (USALEEP) project.[19] Readers should refer to the aforementioned documents for a detailed review of the technical issues related to using GIS software programs. The Tracking Program Standards Network and Development (SND) Geospatial Team consulted with these organizations when working on the Geocoding Standards document. The Geocoding Standards incorporated the lessons learned from the work of others and is reflective of the needs of the Tracking Program to geocode to a sub-county level.

Geocoding Pilots

Sub-County Data Pilot Project (2014–2015)

To move toward working with sub-county data, the Tracking Program proposed the Sub-County Data Pilot Project in 2014, allowing recipients to apply for funding. This pilot projected revealed opportunities and challenges around the collection, transformation, and display of data at a sub-county level. Participating recipients were allowed to select from nearly any relevant health or environmental datasets (e.g., private well water data, acute myocardial infarction hospitalization data). They were also allowed to independently choose the geographic units (e.g., ZIP codes, census tract) and temporal aggregation of the data submitted. From the pilot project, the Tracking Program learned about various concerns, including inconsistent geographic levels, and issues involving aggregation, geocoding data, and data stability.[20]

The resources required to manage and organize inconsistent geographic units summarized across different years was not a sustainable option for the Tracking Program. Additionally, the need to display these data in a useful and nationally consistent way was nearly impossible. The Sub-County Data Pilot Project resulted in a recommendation to develop standardized sub-county geographies using census tracts as the foundation over a consistent period, where possible.[20] This recommendation is not without limitations. For example, some recipients do not have the data to geocode certain relevant measures to a census tract because they might only receive ZIP code-level data. Standardized geographies allow the Tracking Program to compare data across time, space, and measures. However, some measures might be better displayed using different geographies, making the data more understandable and relevant to end users.

Multiple alternatives to census tracts were considered after the pilot was completed. Some states have well-established sub-county geographies for which data are collected and could be displayed (e.g., Maine's townships). However, none of these state geographies are nationally consistent or stable when compared to census tracts. ZIP codes lack consistency and often change multiple times within a single year. Metropolitan Statistical Areas (MSAs) were reviewed, but they lack complete national coverage and are only classified at the county level. Public Use Microdata Areas (PUMAs) have consistency, but the boundaries often resolve to counties or aggregations of counties to meet PUMA standards, which defeats the goal of increasing the availability and accessibility of sub-county data.

During this pilot, other issues were discussed but were not resolved. These included how to handle post office boxes (P.O. Boxes), relevancy of census tracts to end-users, choice of population denominators, suppression for privacy and security, and records that fail to geocode or geocode with low reliability. Although ZIP codes are typically familiar to the public, most people do not identify with their census tract, let alone their census tract in a given aggregation. However, most people also would not recognize their ZIP code boundary on a map. Showing familiar geographic markers can help to overcome this challenge, including county boundaries, major roads, waterbodies, and other landmarks. Another way to address this challenge is to allow users to search for a location on a map and display data for the corresponding census tract.

SND Geospatial Team Geocoding Pilot (2018–2019)

In 2018, the Standards and Network Development (SND) Geospatial Team created a draft geocoding standard document and requested feedback from recipients. In this draft, two categories were proposed to describe the accuracy of geocoding results. One category would be for results considered high precision and one for results considered low precision. Determining which records would be considered high precision versus low precision was a key part of this discussion. The use of two categories accommodates for those records that might geocode

the accuracy of their assignment to a census tract is less certain. Many recipients were concerned about losing these records and the effect their loss would have on the overall counts and rates, particularly in rural areas where the incidence of a disease might already be low. Concerns about rural routes, match rates, match scores, and P.O. Boxes were also highlighted during work to draft geocoding standards.

The SND Geospatial Team conducted a pilot study to gather quantitative data on these issues. In the pilot, recipients were asked to provide the total number of records geocoded, along with the number of records categorized by potential issues, to better understand the percent of records that fit within each category. After presenting the results from this pilot, the SND Geospatial Team collectively determined how to operationally define the high precision and low precision categories.

Six Tracking Program recipients participated in the geocoding pilot (Colorado, Florida, Missouri, New Hampshire, Utah, and Vermont). The datasets that were geocoded included birth records, death records, home addresses of children who received a blood lead level test, hospitalizations and emergency department visits, and registered businesses. The years of data ranged from 1 year to 7 years. Results indicated that 1) data quality improved over time, with more recent years geocoding more successfully than older years; 2) the match rate varied by dataset, with vital records data providing the highest match rate; and 3) P.O. Boxes and rural routes represented only a small percentage of records (between 0% and 2%). Tracking Program staff and contractors presented the results to recipients at a workshop in Atlanta in May of 2018. As a follow up-to these results, the SND Geospatial Team voted on a number of decision points.

The following were key questions and decision points from the SND Geospatial Team geocoding pilot:

- For a match score from ArcGIS and Texas A&M software, should high precision be defined as 99 or greater or 85 or greater?
    - The team decided high precision would be defined as 85 or greater.
- Should an address that geocodes to a ZIP+4 resolution be considered high precision?
    - The team decided that ZIP+4 records would be defined as high precision.
- Should P.O. boxes be removed entirely or should they be imputed across the census tracts that make up the ZIP code in which the P.O. Box is located?
    - The team recommended imputing, but it decided to allow for the removal of these records as long as the number of records removed is noted in the metadata for the year of the dataset being displayed.
- Should rural routes be removed or processed according to the standard like all other records?
    - The team voted to keep rural routes and process them according to the standard.

Sub-County Content Workgroup Team Geocoding Pilot (2018–2019)

In Fall 2018, recipients from Arizona, Florida, Missouri, New York State, Utah, and Washington, along with CDC staff members, pilot tested selected datasets, including birth (and birth defects), crash fatalities, hospitalizations, and mortality, using a variety of geocoding software tools. The aim of this pilot project was to test the Geocoding Standards and to provide feedback on the process. Participants provided input on the understandability and ease of the process and the number of records that geocoded to the different levels of precision. As in the previous pilot, P.O. Boxes and rural routes comprised a very small percentage of the addresses. Some issues were raised for further clarification in the Geocoding Standards, including certain Centrus/MapMarker codes, how to handle previously geocoded data, and inclusion of out-of-state records in

counts. Overall, the participants reported that the process was appropriate, the levels of precision were acceptable, and they were able to produce the information needed for the metadata.

Final Geocoding Decisions

Issues that arose during the work conducted by the SND Geospatial Team and the Sub-County Content Workgroup include the following.

Choice of Software and Underlying Databases

Geocoding accuracy can be influenced by the software used and the underlying database. A newer database might be more accurate at assigning an address to a point than is an older database. The Tracking Program identified the software programs used most frequently by recipients and provided guidelines for those specific software tools (Centrus/MapMarker, ArcGIS, and Texas A&M Geocoder). The Tracking Program does not provide databases for geocoding. Some recipients identified proprietary software and databases (e.g., E911) they used that were state-specific.

Final decision: Request that metadata include the name of the software used in the geocoding process, the version number, and the underlying databases.

Choice of Geographic Unit

Many sub-county geographies were considered in writing the Geocoding Standards. Recipients suggested MSAs, PUMAs, ZIP codes or ZIP code tabulation areas (ZCTAs), and state-specific regions (e.g., towns); however, some of these were not feasible for the Tracking Program's needs. Census tracts were selected as the best sub-county geography because they are relatively stable over time, have well-defined boundaries (unlike ZIP codes), are nationally consistent (unlike state-specific regions), and provide a finer spatial resolution for data analysis than do other sub-county geographies. The Tracking Program recognizes that use of higher resolution geographic units raises concerns related to privacy and stability. The Tracking Program created aggregations of census tracts to use where census tract counts cannot be presented on the public portal because of too few counts, data suppression, or unstable rates.

Final decision: Use census tracts for the sub-county geography (or aggregations of census tracts where there are too few counts, data suppression, or unstable rates).

Use of Dichotomous Categorization

Geocoding "success" depends on many factors.[2, 11, 21, 22] One way in which differences in geocoding success rates systematically differ is by whether an address is rural or urban. Rural addresses are more difficult to geocode to a fine spatial resolution (e.g., census tracts) than are urban addresses. The SND Geospatial Team discussed how to manage the balance between geocoding precision/accuracy and address inclusion.

Briefly, there is an inverse relationship between the number of addresses that will geocode to a point "successfully" and the strength of the definition for that "successful" geocode. Recipients felt very strongly that a case count is still an event, even if it cannot be assigned to a single point with a high level of precision. That is, to exclude all "failed" geocodes in order to support having only high-quality matches was detrimental to having accurate counts and rates. If the Tracking Program created standards that were too strict, there would likely be a lower number of high-quality matches, with rural addresses likely to be disproportionately "failed" at the

census tract level. By excluding case counts from small or rural census tracts, bias will be introduced and the rates for the affected areas might be distorted.

To manage the balance between having accurate counts and presenting quality data, the Tracking Program followed the example of other agencies in using a tiered system for geocoding quality. This would allow for identifying each geocoded address as high precision, low precision, county only, state only, or unknown. Dichotomizing census tract-level geocoding success into high precision and low precision categories will allow the end-user to make judgements about the best data that are suitable for an analysis. The high/low precision categorization scheme allows for counting the maximum amount of addresses while presenting high-quality data. This approach is similar to the approach used by the NCHS USALEEP study.[19]

High Precision

High precision is the aggregation of records being reported to the Tracking Program that clearly meet the minimum requirements for precise and accurate geocoding.

Low Precision

Recipients expressed concern about a loss of certain records due to geocoding to a census tract boundary level. To preserve those records and keep rates from being affected greatly, the Tracking Program decided to allow recipients to also report matches with lower confidence in geocoding precision and accuracy.

The low precision records will mainly include data that matched to a census tract or a geography within a census tract but did not meet the minimum match score. A minimum match score for low precision was not assigned because the methods of generating these scores is not published and the minimum match score and relevant ranges could not be determined. Another type of record expected to be in the low precision grouping are data that matched to larger geographic boundary but can be reasonably placed into the correct census tract or a census tract that represents a proper proportion of the population of the larger geography through imputation. The method and standards for this type of imputation should be determined by recipients. If recipients cannot reasonably place a record in the low precision grouping, it is expected that they will place the record in a county or state aggregation grouping.

The low precision grouping intentionally provides recipients with the ability to determine what data should be included. This allows recipients to use their knowledge of their local populations, geography, and software tools to make decisions about how their records should be treated.

The Tracking Program is still determining how to communicate to end users how the results that include low precision geocodes should be treated, as compared with results that only include high precision geocodes. One idea for the display of census tracts is to flag census tracts with a high proportion of low precision geocodes. Ultimately, the dichotomous categorization scheme and data display will provide users a greater understanding of the geocoding accuracy and precision.

Final decision: Use a dichotomous categorization scheme to include high precision and low precision so each geocoded address is placed into one of these categories.

Use of Geographic Imputation

Geographic imputation refers to deterministic or stochastic methods used to assign a previously non-geocoded record to a “reasonable” location based on other available demographic and spatial information. The Tracking Program had two decisions to make regarding geographic imputation. These were 1) whether to accept

geographic imputation as part of the Geocoding Standards and 2) whether to issue recommendations around geographic imputation methods used by recipients. Recipients shared that some agencies used imputation regularly and others did not. The available literature on geographic imputation suggests accuracy depends largely on the context and specific methods.

Final decision: Include geographic imputation in the Geocoding Standards and have recipients use their state's standard geographic imputation methods.

Minimum Match Score

As mentioned previously, even though Texas A&M Geocoder and ArcGIS software both output a match score for a given address, the match scores are not equivalent. Match score is not a measure of precision; however, match score is the best available proxy measure of geocoding quality. Recipients provided input regarding the best match score to differentiate between high and low precision at the census tract level. The SND Geospatial Team pilot (referred to in the previous section on Geocoding Pilots) identified how different match scores affected the number of high precision and low precision counts.

The match score will depend on a number of factors, including the target spatial resolution of the end product, the quality of the input address data, and the type and quality of the reference address database used by the specific geocoding software. Thus, the use of a minimum match score is only meaningful insofar as it applies to a geocoding result at a specific geographic resolution (e.g., residential address, census block, tract, or county). Because the goal of the Tracking Program's Geocoding Standards is to provide guidelines for transforming address-level health data to the census tract level, it is recommended that a match score of 85 or greater be classified as high precision, provided the result matched to a location (e.g., point, street segment, block, ZIP code, etc.) contained within a census tract.

Final decision: Use a match score of 85 or greater for the high precision category.

Treatment of Group Quarters

Another issue was whether to include or exclude group quarters (e.g., college dorms or prisons)[23] in the census tract counts. Some recipients had databases already in place to identify group quarters, but others did not. Census tracts that include group quarters might have artificially high rates due to the inclusion of a large number of people not actually identified as a usual resident of that census tract. However, several points led to the decision to include group quarters (i.e., do not treat them any differently).

The American Communities Survey includes institutional and non-institutional quarters, with the sample including anyone residing for at least 2 months at a given address, and the decennial census includes institutional and non-institutional quarters counting residents at their usual address. Populations are typically included in the numerator data (e.g., emergency department discharge data). Based on this, it is plausible to assume most people in group quarters would be captured in the numerator and denominator. An exception is where there are institutionalized populations with their own facility (e.g., prison with a healthcare facility).

Final decision: Treat group quarters the same as any other counts and do not remove from the geocoding process.

Treatment of P.O. Boxes

Studies suggest that P.O. Box addresses cannot be reliably geocoded directly to an appropriate census tract because they often do not match up with the location of a residence.[10, 12, 24] This might be especially true in rural jurisdictions where P.O. Boxes can cover a large geographic area. However, the effect of removing P.O. Boxes can vary among regions, with some jurisdictions disproportionately affected by the exclusion of P.O. Box addresses while others experience little effect. Thus, the Geocoding Standards give recipients flexibility in how they address the issue of P.O. Boxes. Recipients are asked to record their process for addressing P.O. Boxes in the metadata. The Geocoding Standards do not recommend geocoding P.O. Boxes directly to census tracts. Grantees may impute to a census tract based on a P.O. Box address falling within an alternate sub-county unit (e.g., town or 5-digit ZIP code) or remove them if the percentage of total records is deemed inconsequential by the recipient. All P.O. Box records will continue to be counted at the county-and state-level.

Final decision: Record the process for assigning P.O. Boxes in the metadata and do not geocode P.O. Boxes directly to census tracts.

Treatment of Rural Routes

As with P.O. Box addresses, rural route addresses used for mail delivery are not accurate reflections of the exact physical location of a residence. However, unlike P.O. Boxes, it might be more feasible to use standard geocoding software with address conversion capability or E911 readdressing methods to recover and geocode valid street addresses from rural route delivery boxes.[25] It also might be possible to assign a rural route address to a census tract with a reasonable level of accuracy if the entire rural route is contained within a census tract, especially in rural areas where census tracts tend to cover a larger geographic area.

Final decision: Treat rural route addresses the same as a standard street address and geocoded via the standard process.

Choice of Boundary Year(s)

Census tract boundaries change over time, usually with changing of the decennial census. After selecting census tracts as the sub-county unit for which the Tracking Program would receive counts, the SND Geospatial Team

discussed which boundaries files would be appropriate. The original proposal suggested by group members was that all health outcome counts should be placed in the census tracts of the boundaries related to the previous decennial census. Counts from 2000–2009 would be placed according to the 2000 Census, counts from 2010–2019 would be placed in boundaries from the 2010 Census, and so on. Placing counts into the boundaries by decennial census creates a problem where comparisons over multiple decades will not have the same boundary shapes. Other solutions that were suggested were to geocode all records to 2010 (the current decennial census) and then re-geocode all records to the Census 2020 boundaries once they become available. Because the decennial census provides boundaries and population counts related to those polygons, some group members felt that calculating rates using population counts from decades past would not provide an accurate rate for some years. Additionally, the burden of re-geocoding counts to different boundaries every 10 years was a concern.

Final decision: Geocode everything 2019 and earlier to 2010 boundaries. Then, all data from 2020-2029 are geocoded to 2020 boundaries (and so on for each subsequent decade).

Use of Centrus/MapMarker Codes

Despite its limitations, the Geocoding Standards recommends using the match score as a basic measure of minimum geocoding quality. ArcGIS and Texas A&M products both produce the match score as standard output, although the match score is not equivalent between the two programs. A third, widely used geocoding software, Centrus/MapMarker, does not produce match scores. This made it necessary to provide guidance on how to produce a basic level of agreement between Centrus/MapMarker and other platforms.

Centrus/MapMarker provides the same three primary outputs for evaluating geocoding quality: result, match, and location codes. Result codes capture which address components (e.g., house number, street prefix/suffix, street name) matched between input data and the geocoding reference data in the form of an eight-character text code, where each character can take one of two values. Like result codes, match codes indicate which address components matched between input and reference data by summarizing which address components in the input data needed to be changed by the match algorithm to create a match to the reference data. For example, such changes often occur because certain address components might be missing in the input data (e.g., street prefix/suffix, street type). Location codes give an indication of the locational precision of the assigned geocode using a three- or four-character alphanumeric code. For example, the location code can be used to distinguish between a point level address geocoding result and an area centroid geocoding result.

The main decision point for the team involved choosing which Centrus/MapMarker output was best to use for establishing an approximate equivalent to output from other software based on match score. To inform this decision, recipients who use Centrus/Mapmarker participated in two geocoding studies. In an initial pilot, the Colorado Tracking Program geocoded 27,832 records from a full year of lead testing data. In a second, larger study, the Arizona Tracking Program geocoded approximately 1.6 million hospital discharges reported to the Arizona Department of Health Services in the first half of 2018. The results of the studies were consistent and suggested that a relatively simpler set of location codes indicating an address, block group, or census tract-level geocoding precision were also generally associated with result and match codes indicating a mismatch of two or fewer address components. Based on these results and given the relative complexity of selecting a comparable set of result or match codes, the team developed a Centrus/Mapmarker equivalent based on a simplified selection of location codes.

One concern about the selected location codes was that it might place a stricter standard on high precision geocoding results from grantees using Centrus/MapMarker software. This is because Centrus/MapMarker has a

more detail-rich output that might allow for greater sensitivity to distinguish the quality of geocoding results than output based only on match score. The team ultimately decided to accept this as a possible limitation that could be addressed by future revision.

Final decision: Use the Centrus/MapMarker equivalent, as outlined in the Geocoding Standards, based on a simplified set of location codes.

Default Program Settings

The Tracking Program assumes that entities who are using geocoding software will use the default settings of the software. If the settings are changed, the Tracking Program recommended only allowing for a stricter matching process.

Metadata

Geocoding is a variable and multistep process. The Tracking Program provided guidance on a standardized method that acknowledges the balance between accepting high quality data and including the maximum health outcome counts at the census tract level. The Tracking Program recommended thoroughly documenting any geocoding process. The Tracking Program recommended that recipients and partners report a specified list of information as metadata, at a minimum. The metadata list includes who went through the process of geocoding, if the data were geocoded using the Tracking Program Geocoding Standards, what underlying reference databases were used, how many records with P.O. Box addresses were in the input dataset, etc. The full list of metadata questions is included in the Tracking Program Geocoding Standards document.

Sometimes data arrive geocoded to the census tract by other entities (e.g., a data steward or registry). The Tracking Program recommended discussing how data were geocoded with the data supplier. The sub-county metadata questions related to geocoding can serve as a useful resource to guide conversations between data receivers and the persons who conduct geocoding (e.g., data steward).

Conclusion

The aim of the 2017–2018 work for the SND Geospatial Team was to contribute to the Tracking Program's sub-county efforts. The specific goal was to develop a set of geocoding guidelines for recipients and partners to use when geocoding address-level data to the census tract. The SND Geospatial Team successfully piloted the geocoding questions that contributed to the Geocoding Standards and received additional feedback on the final Geocoding Standards from a second pilot through the Sub-County Content Workgroup. Many issues were raised and addressed during the team's work, including setting a minimum match score, handling of P.O. Boxes, and imputation, among others. The Geocoding Standards will be put into practice and the Tracking Program will include a geocoding precision variable in future sub-county data calls.

Acknowledgments

The Tracking Program thanks Samantha Wotiz, Angie Werner, and Craig Kassinger (CDC Tracking Program) and Kevin Berg (Colorado Tracking Program) for compiling the information in this document. The Tracking Program also thanks Dan Goldberg from Texas A&M, Recinda Sherman from NAACCR, and Stephanie Foster from the Geospatial Research, Analysis, and Services Program (GRASP) at CDC for providing expertise and consultation as the Geocoding Rationale and Standards documents were developed.

The authors thank the following members of the SND Geospatial Team and the Sub-County Content Workgroup for their efforts and input:

Arizona Department of Health Services: Josue Barboza, Nicole Eiden, David Olsen, Wes Kortuem, Marla Kostuk; **California Department of Public Health**: Heather Amato, Jhaqueline Valle; **Centers for Disease Control and Prevention**: Craig Kassinger, Aaron Grober, Bob Kennedy, Michele Monti, TJ Pierce, Emily Prezzato, Meekie Shin, Jen Shriber, Heather Strosnider, Aaron Vinson, Patrick Wall, Angela Werner, Sam Wotiz; **Colorado Department of Public Health & Environment**: Kevin Berg, Ben White, Devon Williford; **Connecticut Department of Public Health**: Gary Archambault, Patricia Przysiecki; **Delaware Health and Social Services**: Tabatha Offutt-Powell; **Florida Department of Health**: Chris DuClos, John Folsom, Jessi Joiner, Melissa Jordan, Keshia Reid; **Green River**: Michael Knapp; **Kansas Department of Health and Environment**: Jaime Gabel, Henri Menager; **Louisiana Department of Health**: John Anderson, Kathleen Aubin, Kate Friedman, Adrian Savella, Alexis Williams; **Massachusetts Department of Public Health**: Glennon Beresin, Erin Collins, Braden Miller; **Maine Department of Health and Human Services**: Jessica Bonthius, Kathy Decker, Rebecca Lincoln, Lisa Parker, Chris Paulu; **Michigan Department of Health & Human Services**:: Jill Maras, Sydney Ogden; **Minnesota Department of Health**: Tess Konen, Blair Sevcik, Jessie Shmool; **Missouri Department of Health and Senior Services**: Kathleen Kloeppel, Jeff Patridge, Scott Patterson, Elizabeth Semkiw, Jen Weaver; **NAACCR**: Recinda Sherman; **NAHDO**: Charles Hawley, Denise Love, Emily Sullivan; **NAPHSIS**: Raquel Brown, Andrea Price, Kristin Simpson, Shae Sutton; **New Hampshire Department of Health & Human Services**: Katie Bush, Samuel Harris, Dennis Holt, Jennifer Howley, Jessica Sagona, Nicholas Shonka; **New Jersey Department of Health**: Barb Goun, Rick Opiekun; **New Mexico Department of Health**: Tony Fristachi, Lois Haggard, Brian Woods; **New York City Department of Health and Mental Hygiene**: Grant Pezeshki; **New York State Department of Health**: Jeff Bryant, Douglas Done, Tabassum Insaf, Sanjaya Kumar, Neil Muscatiello, Seema Nayak, Arjita Raj, Abby Stamm; **Oregon Health Authority**: Eric Main; **Rhode Island Department of Health**: Peter DiPippo, Joseph Maya-Rodriguez, Jay Metzger, Catherine Schultz, Mike Simoli; **Ross Strategic**: Mary Byrne, Jessie Doody, Samer Khan, Jen Major, Lissette Palestro; **Texas A&M**: Payton Baldridge, Dan Goldberg; **Utah Department of Health**: Johnny Auld, Sam LeFevre, Nelson Long, Matt McCord; **Vermont Department of Health**: Daniel Jarvis, Pete Young; **Washington State Department of Health**: Chris Ahmed, Lauren Freelander, Buffi LaDue, Lillian Morris; **Wisconsin Department of Health Services**: Jenny Camponeschi, Paul Creswell.

References

1. Zimmerman DL, Fang X, Mazumdar S, Rushton G. Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics. 2007;6(1):1-16.

2. Goldberg DW, Swift JN, Wilson JP. Geocoding best practices: Analysis of geocoding requirements. Los Angeles, CA: University of Southern California GIS Research Laboratory, 2008. Technical Report No. 9.

3. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, et al. Geocoding in cancer research: A review. American Journal of Preventive Medicine. 2006;30(2, Supplement):S16-S24.

4. Zhan FB, Brender JD, Lima ID, Suarez L, Langlois PH. Match rate and positional accuracy of two geocoding methods for epidemiologic research. Annals of Epidemiology. 2006;16(11):842-9.

5. Centrus Desktop. User Guide for Windows. Pitney Bows Software Inc.; 2018.

6. Map Marker Plus. User Guide. Pitney Bows Software, Inc.; 2017.

7. ESRI. About modifying an address locator's settings. 2018 [cited 2018 July 11]. Available from: https://desktop.arcgis.com/en/arcmap/latest/manage-data/geocoding/modifying-an-address-locator-s-settings-about.htm.

8. ESRI. Geocoding: Delivering high location accuracy. 2017 [cited 2018 July 11]. Available from: https://www.esri.com/arcgis-blog/products/arcgis-enterprise/analytics/geocoding-delivering-high-location-accuracy/.

9. Ha S, Hu H, Mao L, Roussos-Ross D, Roth J, Xu X. Potential selection bias associated with using geocoded birth records for epidemiologic research. Annals of Epidemiology. 2016;26(3):204-11.

10. Zandbergen PA. Geocoding quality and implications for spatial analysis. Geography Compass. 2009;3(2):647-80.

11. Baker J, Alcantara A, Ruan X, Watkins K. The impact of incomplete geocoding on small area population estimates. Journal of Population Research. 2012;29(1):91-112.

12. Oliver MN, Matthews KA, Siadaty M, Hauck FR, Pickle LW. Geographic bias related to geocoding in epidemiologic studies. International Journal of Health Geographics. 2005;4(1):1-9.

13. Henry KA, Boscoe FP. Estimating the accuracy of geographical imputation. International Journal of Health Geographics. 2008;7:3.

14. Dilekli N, Janitz AE, Campbell JE, de Beurs KM. Evaluation of geoimputation strategies in a large case study. International Journal of Health Geographics. 2018;17(30).

15. Baker J, White N, Mengersen K. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. International Journal of Health Geographics. 2014;13(1):47.

16. Hibbert JD, Liese AD, Lawson A, Porter DE, Puett RC, Standiford D, et al. Evaluating geographic imputation approaches for zip code level data: An application to a study of pediatric diabetes. International Journal of Health Geographics. 2009;8:54.

17. Jones SG, Ashby AJ, Momin SR, Naidoo A. Spatial implications associated with using Euclidean distance measurements and geographic centroid imputation in health care research. Health Services Research. 2010;45(1):316-27.

18. Swift J, Goldberg D, Wilson J. Geocoding best practices: Review of eight commonly used geocoding systems. Los Angeles, CA: University of Southern California GIS Research Laboratory. 2008.

19. Arias E, Escobedo LA, Kennedy J, Fu C, Cisewki J. U.S. Small-area Life Expectancy Estimates Project: Methodology and results summary. Vital Health Statistics. 2018 Sep(181):1-40.

20. Werner AK, Strosnider H, Kassinger C, Shin M. Lessons learned from the Environmental Public Health Tracking sub-county data pilot project. Journal of Public Health Management and Practice. 2018;24(5):E20-E7.

21. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics. 2003;2:10.

22. Chow TE, Dede-Bamfo N, Dahal KR. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. Annals of GIS. 2016;22(1):29-42.

23. U.S. Census Bureau. Group quarters and residence rules for poverty. 2018 [cited 2018 February 22]. Available from: https://www.census.gov/topics/income-poverty/poverty/guidance/group-quarters.html.

24. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post office box addresses: A challenge for geographic information system-based studies. Epidemiology. 2003:386-91.

25. Vieira VM, Howard GJ, Gallagher LG, Fletcher T. Geocoding rural addresses in a community contaminated by PFOA: A comparison of methods. Environmental Health. 2010;9(1):1-7.