Accepted Manuscript online: 2023-08-04 Article published online: 2023-10-18

POINT: Pipeline for Offline Conversion and Integration of Geocodes and Neighborhood Data

Kevin Guo[1] Allison B. McCoy[2] Thomas J. Reese[2] Adam Wright[2] Samuel Trent Rosenbloom[2] Siru Liu[2] Elise M. Russo[2] Bryan D. Steitz[2]

[1] School of Medicine, Vanderbilt University, Nashville, Tennessee, United States
[2] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, United States

Address for correspondence Bryan D. Steitz, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 1475, Nashville, TN 37203, United States (e-mail: Bryan.d.steitz@vumc.org).

Appl Clin Inform 2023;14:833–842.

Abstract

Objectives Geocoding, the process of converting addresses into precise geographic coordinates, allows researchers and health systems to obtain neighborhood-level estimates of social determinants of health. This information supports opportunities to personalize care and interventions for individual patients based on the environments where they live. We developed an integrated offline geocoding pipeline to streamline the process of obtaining address-based variables, which can be integrated into existing data processing pipelines.

Methods POINT is a web-based, containerized, application for geocoding addresses that can be deployed offline and made available to multiple users across an organization. Our application supports use through both a graphical user interface and application programming interface to query geographic variables, by census tract, without exposing sensitive patient data. We evaluated our application's performance using two datasets: one consisting of 1 million nationally representative addresses sampled from Open Addresses, and the other consisting of 3,096 previously geocoded patient addresses.

Results A total of 99.4 and 99.8% of addresses in the Open Addresses and patient addresses datasets, respectively, were geocoded successfully. Census tract assignment was concordant with reference in greater than 90% of addresses for both datasets. Among successful geocodes, median (interquartile range) distances from reference coordinates were 52.5 (26.5–119.4) and 14.5 (10.9–24.6) m for the two datasets.

Conclusion POINT successfully geocodes more addresses and yields similar accuracy to existing solutions, including the U.S. Census Bureau's official geocoder. Addresses are considered protected health information and cannot be shared with common online geocoding services. POINT is an offline solution that enables scalability to multiple users and integrates downstream mapping to neighborhood-level variables with a pipeline that allows users to incorporate additional datasets as they become available. As health systems and researchers continue to explore and improve health equity, it is essential to quickly and accurately obtain neighborhood variables in a Health Insurance Portability and Accountability Act (HIPAA)-compliant way.

Keywords

- social determinants of health
- geographic mapping
- census tract
- personalized medicine

received
May 2, 2023
accepted after revision
August 3, 2023
accepted manuscript online
August 4, 2023

© 2023. Thieme. All rights reserved.
Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/a-2148-6414.
ISSN 1869-0327.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Background and Significance

The environments in which individuals live, work, and socialize greatly influence health and well-being.[1,2] These factors, known as social determinants of health (SDOH), are upstream from specific disease processes, but influence a person's chances to be healthy.[3,4] SDOH is a key contributor to many health disparities, which are partially responsible for disproportionate trends of morbidity and mortality at a population level.[5–7] Disadvantaging SDOH such as low education, poverty, limited access to health care, and social isolation are associated with both increased risk of developing and having worse outcomes due to disease states such as diabetes, cardiovascular disease, and kidney disease.[8–13] Identifying patients' SDOH can inform research and support patient-specific health care needs and interventions, although it is important to consider issues of data quality, spatial ambiguity, and population fallacy when relying on neighborhood-level estimates.

Despite the importance of SDOH to patient health and well-being, electronic health records (EHRs) seldom capture structured data about SDOH.[14–17] Factors that contribute to this issue include a lack universally agreed-upon SDOH, lack of structured fields within the EHR, and increased workload for health care workers who collect and input these data.[14,16]

One approach to inferring SDOH is to estimate based on where the patient lives.[18] Organizations such as the Agency for Healthcare Research and Quality and the Center for Disease Control and Prevention routinely publish SDOH data delineated by census boundaries.[19–21] Measuring SDOH by census boundaries, most commonly census tract, allows for granular calculations that closely represent the community. Obtaining boundary details requires calculations using addresses that are available in the EHR. There is tremendous heterogeneity in the size and population between ZIP codes.[22] The U.S. Census Bureau defines smaller increments, such as census tracts and block groups, that are more uniform in size and population.[22,23]

Geocoding, or converting addresses into geographical coordinates, allows researchers to obtain neighborhood-level estimates of SDOH.[24] Geocoding is performed via two methods: offline geocoding and geocoding through an online service. Offline geocoding software such as DeGAUSS, Nominatim, EaserGeocoder, SAS Geocoder, ESRI ArcGIS, QGIS, and the PostGIS TIGER geocoder[25–30] have been available for several years, but they often require an expensive license or come with steep learning curves. Online tools, such as Google Maps, require sharing addresses with the service, which risks privacy concerns. Under the Health Insurance Portability and Accountability Act (HIPAA) privacy rule, addresses and census-level data are considered protected health information.[31] Per-address fee structures also prove costly when geocoding large datasets. Both methods require the additional step of mapping from geographic coordinates to SDOH to be performed separately.

Objectives

Despite the importance of SDOH for research and operational use, there remains a critical need for a local, HIPAA-compliant geocoding platform that can be easily deployed across an organization and available to researchers and at the point of care. We developed POINT: an interactive, web-based, containerized, application for geocoding addresses that can be deployed offline and available to multiple users with minimal technical expertise. Our application supports use through both a graphical user interface (GUI) and application programming interface (API) client to query geographic variables, by census tract and across census years, without deploying their own solution or exposing sensitive patient data. Integrating SDOH databases into the geocoding workflow streamlines the process and allows for customizability to fit user needs. POINT serves as a low-cost and scalable alternative to using a web service.

Methods

Technical Design

POINT uses Topographically Integrated Geographic Encoding and Referencing (TIGER) Line files.[32] TIGER/Line files are maintained by the U.S. Census Bureau and contain coordinate boundaries down to street and street number. Every census geographic area is identified by a unique Federal Information Processing Standard (FIPS) code. The Census TIGER/Line files are organized into key components by census county: census-designated places or incorporated places, county subdivisions, census tracts, census block groups, topological faces, names of each line/geographic area, line coordinates, and address ranges, of which a separate set of shape files exist for each county.[32] Each file is downloaded, programmatically transformed, and imported in to a PostgreSQL database for address-level mapping.[33]

An overview of the system architecture is displayed in ►Fig. 1. We loaded census boundaries into PostGIS, a geographic information system (GIS) enabled database, to support address-level mapping into geographic coordinates, which are then converted into census boundaries using structured query language (SQL). PostGIS, the spatial database extension for PostgreSQL, provides robust functionality to standardize and geocode address strings.[28] The address standardization process involves regular expression to determine the type of address, identify address components (such as ZIP code or street name), and parse the address into a standard data structure with each component clearly delineated. Our geocoder platform and supporting files are available on our GitHub repository.[34]

To package our software, we created a containerized system consisting of two images: one for the database and one for our Python-based uvicorn web server.[35] We deploy the containers using Docker, a virtualization platform that facilitates portability and reproducibility across systems and organizations.[36] A python script is included that assists with the process of importing data from common SDOH databases, including PLACES: Local Data for Better Health, Agency for Healthcare Research and Quality SDOH Database, CDC Social Vulnerability Index, Food Environment Atlas, Community Resilience Estimates, Area Deprivation Index, and United States Department of Agriculture

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

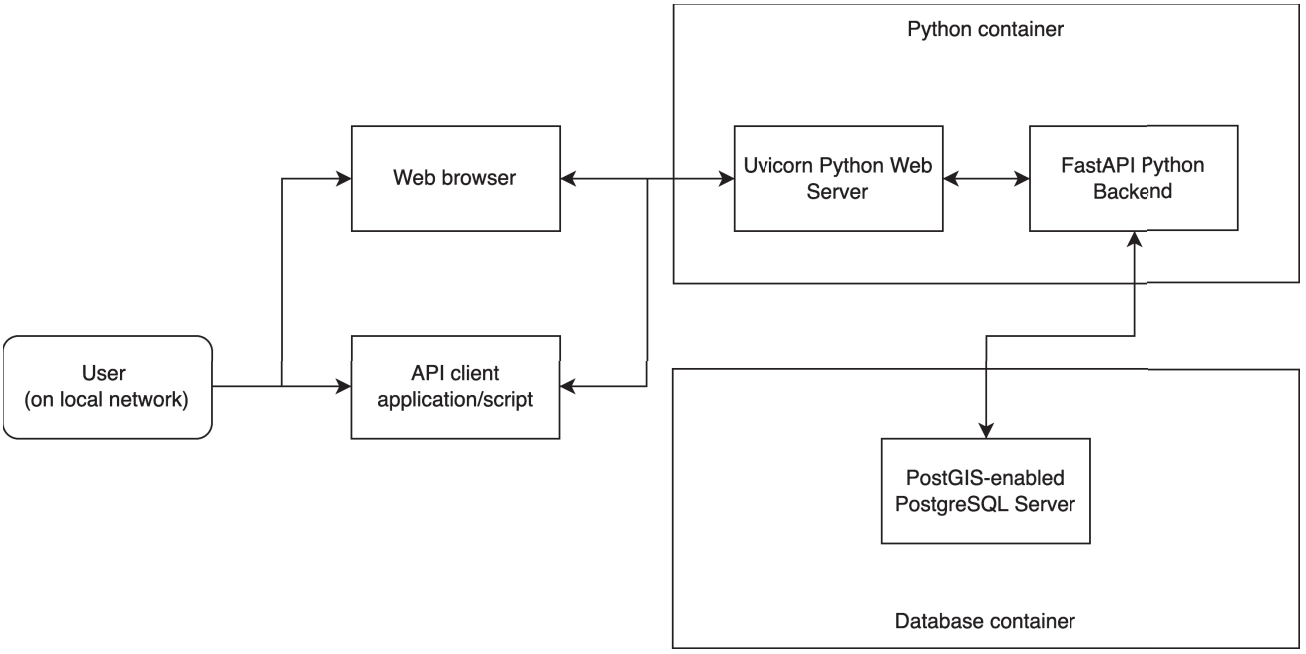

Fig. 1 System architecture diagram showing interactions between each component of the application.

rural–urban communicating areas (RUCA).[19,21,37–42] Users may add SDOH mappings using an included Python script that loads a character-delimited file with variable values for each FIPS code (county, tract, or block group). Future census boundaries, or other types of spatial data (such as Health Resources and Services Administration shortage areas), can be imported by running the included PostGIS functions or importing the shape (.shp) file(s).

To support multiple users, each geocoding job, by default, is identified with an integer number and password to maintain privacy and security when processing sensitive patient data. The password and job number are required to access results. Some organizations implementing POINT may wish to disable this feature and integrate the tool with local security resources. Addresses and a user-defined identifier are saved in the database for the duration of the geocoding job and are deleted automatically 72 hours after job completion or 1 hour after download. Temporary files are generated during a download, then immediately deleted.

We made a docker-compose configuration file that gives docker instructions to pull the docker images from the docker repository and deploy the application. Also included is a shell script that will load the full 2021 Census TIGER dataset along with 2010 block group boundaries. It can take several hours to download and import all data, depending on the system and download speeds, and will take up about 100 to 130 GB of disk space. To reduce disk space, scripts are available to download data for only a subset of states.

User Interface and System Functionality

The user interface supports two modes of access: a GUI and an API to support programmatic queries. The API conforms to representational state transfer architecture and OpenAPI specifications.[43] Our platform provides both geocoding (converting address to Census FIPS code) and geovariable mapping (looking up values of variables based on Census FIPS code) functions that can be performed either together or separately (►Fig. 2).

To geocode, the user inputs a character delimited file with columns corresponding to an address or individual address fields. The application outputs coordinates (longitude/latitude), census block groups FIPS codes, and geocoding scores (0–138 estimate of geocoding accuracy/resolution [0 being an exact match]). Based on our experiments, we set the default threshold for successful geocoding to a rating of 25 or below, but thresholds may be adjusted for different geographic precision. After an input file is uploaded, the user defines a password, and a "job" is created with a unique identifier so that the user can return to check progress or download results.

The web application provides support to map geocoded addresses to a list of geographic variables. The user can select geographic variables from a list of available measures, based on the SDOH sources loaded into the application database. Target geographic variables can be selected prior to geocoding as a part of the batch geocoding job creation workflow or using files that were previously geocoded (►Fig. 2).

System Evaluation

The PostGIS Tiger Geocoder was previously validated against a subset of the Open Addresses dataset[44] using the bench4gis geocoding benchmarking framework with a reported 99% hit rate (successful geocoding to geographic coordinates) and 65% accuracy within 100 m and 90% accuracy within 1 mile.[45]

We evaluated our application's performance using two address datasets. The first dataset contained 1,000,000 nationally representative addresses sampled from Open Addresses, which we have published online.[46] Open Addresses is a public database of street addresses and reference coordinates collected from authoritative sources such

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

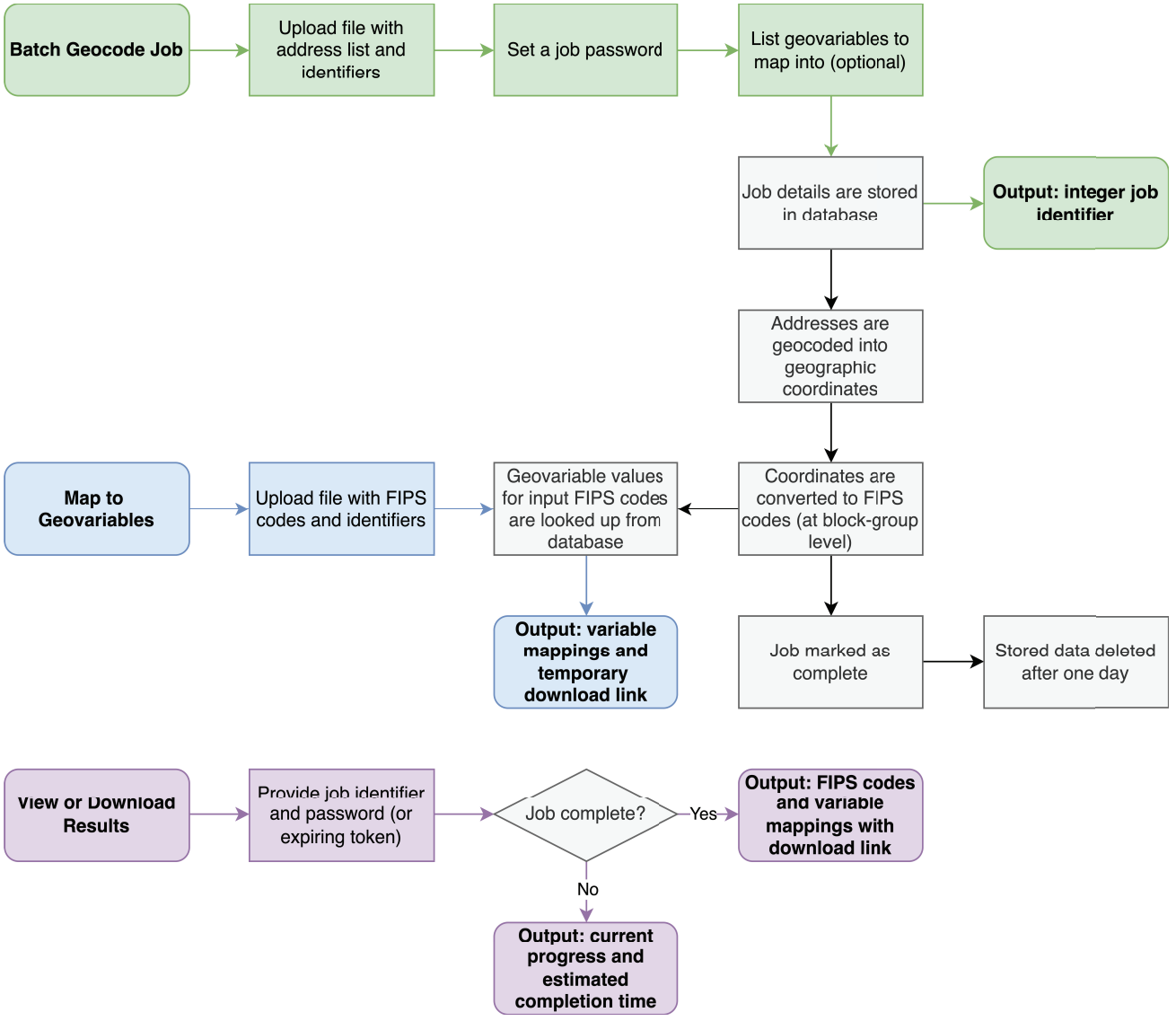

Fig. 2 Workflow diagram of core functionality. Each user workflow (initiating a batch geocode job, mapping geovariables, and viewing/downloading results) is highlighted in a different color. Nodes with gray background represent system processes.

as local GIS departments or postal services.[44] Random addresses were sampled in a population-weighted manner such that the distribution of states in the dataset would match state populations as of the 2020 census. The second dataset contained 3,096 patient addresses from Vanderbilt University Medical Center (VUMC) that were previously geocoded with the official Census.gov geocoder, which we took as gold standard.[47]

First, we geocoded both datasets with the POINT geocoder and the DeGAUSS geocoder to evaluate overall hit rate as a function of rating. For the Open Address dataset, we used our platform's multithreading feature to improve efficiency (4 threads). No equivalent feature was available for the DeGAUSS geocoder. To evaluate geocoder accuracy, we compared concordance in assigned census block group, tract, and county between output from the POINT geocoder with reference coordinates. We also evaluated geocoder accuracy as a function of rating. We defined an error by calculating geodesic distances between coordinates returned by the POINT geocoder and reference coordinates. We visualized the difference between calculated geocodes and reference coordinates using choropleth maps generated using the plotly package in Python version 3.9.[48] To compare geocoder accuracy rating cutoffs, we computed planar census tract areas and compared average tract areas between urban and rural tracts, based on RUCA codes 8, 9, or 10.

Results

Our sample of the Open Addresses dataset consisted of 1,000,000 addresses from 49 of 50 U.S. States. Open Addresses does not contain addresses in New Hampshire, so these were not represented in our sample. The VUMC addresses dataset had 3,096 total addresses, consisting of 2,588 (83.5%) addresses from Tennessee, 249 (8.0%) addresses from Kentucky, and 124 (4.0%) addresses from Alabama. ►Table 1 compares geocoding statistics between POINT and DeGAUSS. Compared with DeGAUSS, POINT

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 1 Comparison of geocoder accuracy and runtimes

	POINT geocoder	DeGAUSS
Open addresses		
Successful geocodes (%)	994,146 (99.4)	963,672 (96.4)
Median error, m (IQR)	52.5 (26.5–119.4)	54.7 (29.8–113.5)
Runtime	31 h	103 h
VUMC dataset		
Successful geocodes (%)	3,089 (99.8)	3,058 (98.8)
Median error, m (IQR)	14.5 (10.9–24.6)	15.5 (9.6–30.1)
Runtime	17 min	21 min

Abbreviations: IQR, interquartile range; VUMC, Vanderbilt University Medical Center.

mapped 30,474 more addresses from the Open Addresses dataset in 30% of the time (31 vs. 103 h). Among successful mappings across both geocoders, performance was similar with a median (interquartile range) distance of 52.5 (26.5–119.4) and 14.5 (10.9–24.6) m from reference for the Open Addresses and VUMC datasets, respectively.

Out of the addresses in the VUMC dataset, 2,907 (93.4%), 2,942 (95.0%), and 3,034 (98.0%) were concordant between the Census.gov and POINT results for census block group, tract, and county levels respectively. ►**Table 2** provides a breakdown of accuracy at the census block group, tract, and county levels by RUCA codes for the Open Addresses dataset, where reference coordinates are available. Among successfully geocoded addresses that could be mapped to RUCA codes, POINT geocoded 888,192 (89.4%), 903,256 (90.9%), and 965,955 (97.2%) addresses to the same census block group, tract, and county levels, respectively. We visualize the difference in census tract concordance as a function of county in ►**Fig. 3**. ►**Table 3** provides detailed accuracy metrics for the POINT geocoder across both datasets. Our geocoder achieved the best-possible accuracy rating of 0, which corresponds to an exact match, in 53.7% of addresses across both datasets. A total of 921,992 (91.8%) addresses were successfully geocoded within our default rating cutoff of 25. Similarly, at a rating cutoff of 25, 63.2% of Open Addresses and 72.4% of VUMC addresses were within 500 m of the reference coordinates. Hit rates across levels of geographic precision are available in ►**Supplementary Table S1** (available in the online version). We include comparison of geocodes calculated from DeGAUSS and POINT in ►**Supplementary Table S2** (available in the online version). Among all successfully geocoded addresses, 31,081 (3.1%) from the Open Addresses dataset and 175 (5.7%) from the VUMC address dataset were identified as rural residences. Specific census tract areas and average tract areas by county are included in ►**Supplementary Table S3** (available in the online version).

Discussion

We developed a web-based application to enable offline, HIPAA-compliant, geocoding, and downstream mapping to

Table 2 Geocoding accuracy at various census divisions for Open Addresses dataset by rural–urban communicating area codes

RUCA code	Frequency, *n* (%)	Correct county, *n* (%)	Correct tract, *n* (%)	Correct block group, *n* (%)
Overall	993,614	965,955 (97.2)	903,256 (90.9)	888,192 (89.4)
1 (metropolitan area core)	791,851 (79.7)	778,554 (98.3)	728,920 (92.1)	718,924 (90.8)
2 (high commuting to metropolitan area)	88,423 (8.9)	82,663 (93.5)	77,186 (87.3)	75,225 (85.1)
3 (low commuting to metropolitan area)	6,968 (0.7)	6,434 (92.3)	5,929 (85.1)	5,729 (82.2)
4 (Micropolitan area core)	39,868 (4.0)	37,883 (95.0)	34,754 (87.2)	34,011 (85.3)
5 (high commuting to micropolitan area)	14,242 (1.4)	13,111 (92.1)	12,437 (87.3)	12,006 (84.3)
6 (low commuting to micropolitan area)	2,685 (2.4)	2,373 (88.4)	2,243 (83.5)	2,174 (81.0)
7 (small town core)	18,496 (1.9)	16,868 (91.2)	15,467 (83.6)	14,833 (80.2)
8 (high commuting to small town)	5,126 (0.5)	4,634 (90.4)	4,365 (85.2)	4,217 (82.3)
9 (low commuting to small town)	2,300 (0.2)	2,049 (89.1)	1,969 (85.6)	1,923 (83.6)
10 (rural areas)	23,655 (2.4)	21,386 (90.4)	19,986 (84.5)	19,150 (81.0)

Abbreviation: RUCA, rural–urban communicating area.
Notes: Reference based on published geographic coordinates. A total of 6,386 (0.64%) addresses were excluded due to inability to map to RUCA code.

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

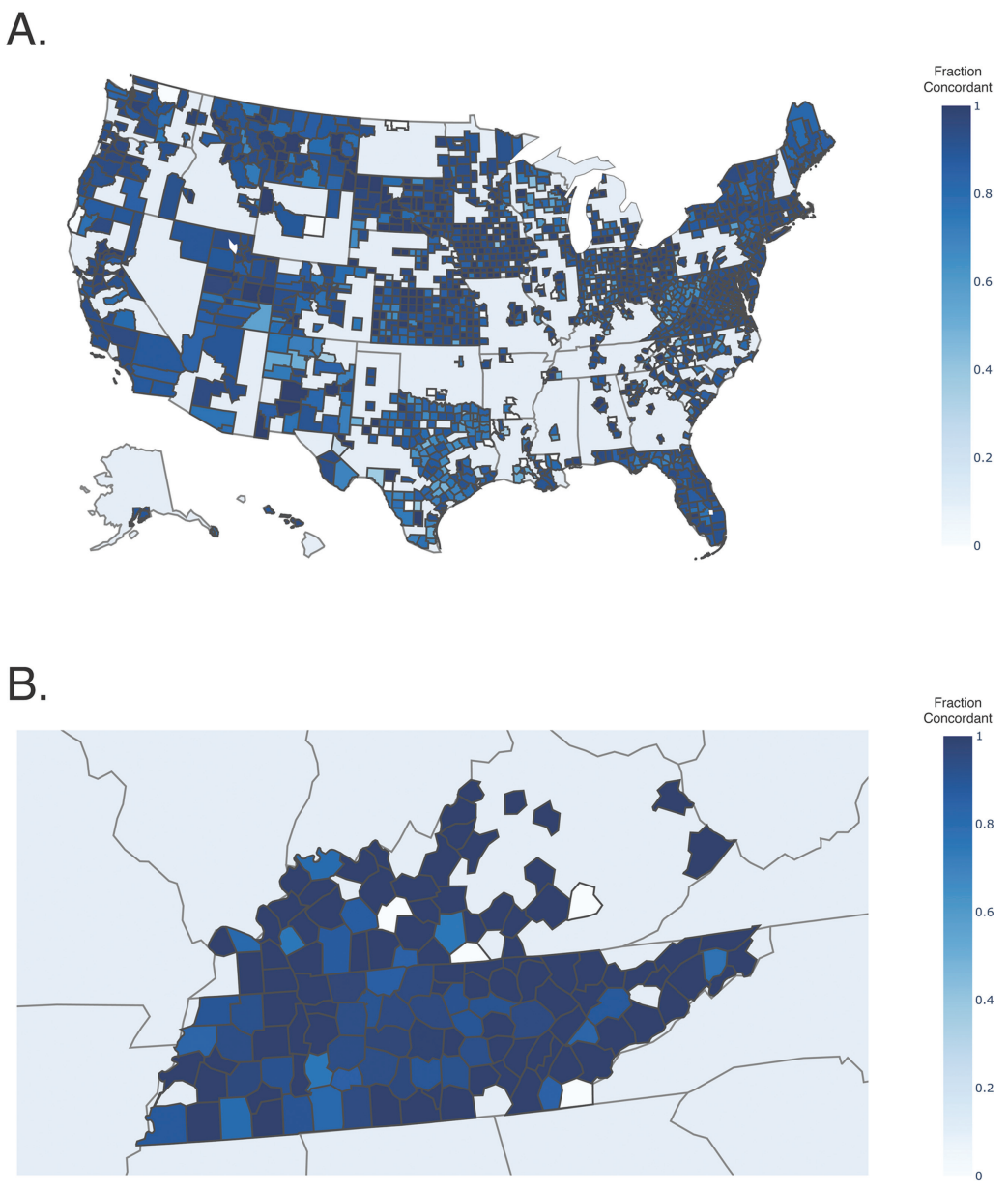

Fig. 3 Choropleth maps showing POINT accuracy at the census tract level (compared with reference) in each county for the (**A**) Open Addresses Dataset and (**B**) VUMC addresses (only Tennessee and Kentucky counties). Counties without at least one address geocoded are indicated in gray. VUMC, Vanderbilt University Medical Center.

neighborhood-level variables. The POINT geocoder includes both a GUI and API to support users across a range of technical expertise. The application supports mapping to multiple census years and sources of neighborhood-level data, and we've integrated a robust pipeline that allows users to incorporate additional datasets as they become available. Our results demonstrate that POINT offers an improved hit rate with similar accuracy to existing solutions, including DeGAUSS and the U.S. Census Bureau's official geocoder.

Understanding community- or neighborhood-level variation is essential to evaluating SDOH and reducing disparity in health and health care.[4,16] For example, community vital signs—aggregate measures of SDOH—have been proposed as a way to integrate community-level social determinants into clinical decision support tools.[18,49] These community vital signs could identify patients who may benefit from targeted interventions, such as sending informational material on quick and easy healthy recipes for patients who live in food deserts. They can also be incorporated into predictive risk modeling at a population level for provider reimbursement adjustments or community-level initiatives.[49–51] Integrating individual patient SDOH into the EHR can support clinical work and improve patient engagement. Using coarsened geocodes such as census division instead of exact patient addresses also serves to preserve individual patient privacy in research.

The POINT geocoder offers several advantages over existing geocoding applications. First, the POINT geocoder was designed to provide free robust geocoding and SDOH mapping capabilities to multiple users across an organization.

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Table 3 Distances in meters between the POINT geocoded coordinates and published Open Addresses coordinates or Census.gov geocoder coordinates for each rating bin

		Distance from reference (m)	Proportion of distances below threshold			
Rating	Addresses, *n* (%)	Median (IQR)	≤50 m	≤100 m	≤500 m	≤1,000 m
Open addresses dataset						
0	538,349 (54.2)	41.9 (23.2–76.8)	57.6	83.0	98.9	99.2
5	75,255 (7.6)	58.1 (28.0–134.3)	44.4	68.0	91.7	95.5
10	165,396 (16.6)	58.2 (29.5–132.8)	44.5	67.8	90.8	93.7
15	86,548 (8.7)	57.3 (28.6–126.8)	44.8	69.0	90.8	93.6
20	35,397 (3.6)	84.2 (35.4–357.9)	34.8	54.4	77.8	82.5
25	18,119 (1.8)	158.0 (49.2–2,571.3)	25.4	41.4	63.2	68.1
50	31,670 (3.2)	4,013.1 (148.9–36,845.5)	12.1	20.7	34.8	39.1
100	40,955 (4.1)	8,039.5 (3008.8–23945.2)	4.2	6.8	11.2	13.7
150	2,457 (0.3)	153,082.5 (12,745.8–293,679.8)	0.0	0.2	1.4	2.8
VUMC addresses dataset						
0	1,095 (35.4)	12.7 (10.4–17.8)	97.1	99.4	99.8	99.8
5	136 (4.4)	19.8 (13.2–57.7)	73.5	79.4	88.2	91.9
10	1,210 (39.2)	14.4 (10.9–23.6)	89.8	94.5	96.6	96.9
15	327 (10.6)	17.5 (11.8–30.5)	85.3	94.8	98.2	98.5
20	102 (3.3)	22.0 (12.8–39.3)	80.4	84.3	89.2	90.2
25	58 (1.9)	22.2 (12.6–1,141.7)	56.9	69	72.4	74.1
50	48 (1.6)	38.9 (13.9–9,757.7)	52.1	54.2	58.3	58.3
100	99 (3.2)	3,420.1 (17.2–63,914.2)	42.2	42.2	43.4	43.4
150	14 (0.5)	7,985.8 (4,898.3–9,689.9)	0	0	0	0

Abbreviations: IQR, interquartile range; VUMC, Vanderbilt University Medical Center.
Note: Percentages reported out of total hits (994,146, and 3,089, respectively).

Existing tools offer free offline services to single users or online services to multiple users. POINT serves as an important intermediate solution between fully offline software packages that each user must configure on their own and an online cloud-based solution that requires exposing sensitive data to a third party. Second, POINT provides access through both GUI and API. Other offline tools often only support a single type of access, most commonly through command line interface. Users with technical expertise can access the tool programmatically and integrate it into established analytic pipelines, whereas users who prefer a graphical interface can perform all tasks through their web browser. At our institution, we are exploring approaches to integrate geocoding into the EHR using the POINT API. One initiative involves geocoding addresses for patients in the emergency department to identify opportunities for convenient follow-up close to home.

POINT provides a single robust pipeline to geocode addresses and map geocodes to SDOH measures. Existing solutions commonly offer geocoding functionality but rely on users to perform additional mapping to SDOH metrics. Providing geocoding and SDOH mapping functionality in a single pipeline supports users without requiring additional technical expertise to curate, transform, and link SDOH data. At our institution, we are experimenting with opportunities to integrate the POINT SDOH pipeline in the EHR as part of decision support to identify patients who may need additional support during telehealth visits. POINT also supports scalability to multiple datasets. By default, POINT incorporates data from the 2010 and 2020 census and multiple commonly referenced SDOH databases. However, census boundaries change every 10 years, and new SDOH datasets are consistently published or updated. POINT includes functionality to import new census years and SDOH datasets.

Our experiments suggest that POINT offers performance that is consistent or superior to existing tools. We were able to corroborate reported benchmark hit rates of 99% with POINT yielding a hit rate of 99.4%.[45] Across census block group, tract, and county, POINT was greater than 93% concordant for addresses in the VUMC dataset. Based on reference coordinates from Open Addresses, we were able to obtain concordant assignments of 89.4, 90.9, and 97.2% at the census block group, tract, and county levels respectively, with expected declines for addresses in areas with decreased population density. Even at the most precise census division (block group), the worst percent concordance was still above 80% in low population density areas. The slightly worse performance for the Open Addresses dataset may reflect lower-quality reference coordinates due to the heterogeneity of address sources in the Open Addresses dataset.

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Concordance between output coordinates suggest that POINT offers similar accuracy to other geocoders (median distances of 14.5 and 5.9 m vs. Census.gov reference and the DeGAUSS geocoder, respectively). We hypothesize that difference in hit rate between POINT and DeGAUSS may reflect differences in prefiltering of poor quality geocodes. On the geographically diverse and nationally representative Open Addresses dataset, concordance with published coordinates was similar with a median distance of 52.5 m. Common reasons for failure include typos in the address string and incorrectly positioned apartment numbers. Future work that advances address string standardization beyond PostGIS functions to better detect and correct typographical errors and ensure consistent formatting prior to geocoding may improve geocoding performance. We recommend that users consider standardizing address strings, such as with a CASS certified software, before using them as input for the POINT geocoder.

Geocoding with online services, such as Google Maps and OpenStreetMaps (OSM), has been evaluated with similar methods.[52,53] Hit rates of 93 and 82% and median distances from reference coordinates of 9 and 175.8 m were previously reported for Google Maps and OSM, respectively.[52] Google Maps yields a slightly better median distance from reference (9 vs. 14.5 m) than POINT. However, the nationwide mean census tract area based on 2020 census boundaries is 116.8 km^2; metropolitan city cores had a median tract area of 8.0 km^2. It is unlikely that the median distance from reference between Google Maps and POINT yields significantly different tract-level results. Use of Google Maps requires exposing addresses to a third-party server.

Spatial uncertainty and data quality are two key considerations in geocoding addresses. One source of spatial uncertainty stems from ambiguous road network data, in that positions for specific street/house numbers are often interpolated based on address ranges when they are not always uniformly distributed across a given street.[54] Our analyses relied on two large datasets that separately provided addresses corresponding to a robust national representation and detailed local representation. However, these datasets suffer from a lack of "ground truth" geocodes and inconsistent data quality. To address this limitation, we assessed concordance between multiple existing geocoders that have been applied widely. In creating our evaluation dataset from Open Addresses, we conducted a weighted sampling approach to sample respective to the population of each state. While this approach yielded a nationally representative sample, county representation in some states was incomplete or poor. This was likely due to how data were collated to create the Open Addresses dataset, which used a large variety of local sources, some of which did not provide complete data or with improperly labeled address segments for inclusion in the evaluation set. Analysis of accuracy may also differ significantly between established and new communities, especially those whose street names are new, and we did not have a good method to systematically identify newer addresses. Finally, a limitation of using overall hit rates is that a reported successful geocode does not necessarily imply accuracy. The PostGIS geocoder, for example, will return successful geocodes at geographic centroids of census-designated places or ZIP codes if street number/name cannot be matched to one in the database. With every geocoding attempt, the PostGIS geocoder returns a rating score based on confidence.[28] While we propose 25 as a potential threshold for accuracy, alternative thresholds may be more appropriate for different datasets, tasks, or research questions. Users may wish to investigate appropriate cutoffs for their specific projects. For instance, geocoding error rate increases as population density decreases.[55] This is an observation that we have redemonstrated in ►Table 2.

There are several limitations to geocoding in research and operational settings. Firstly, it is important to consider the risk of ecological fallacy when using geocoding as a tool to estimate individual patient characteristics. Aggregate SDOH characteristics based on home addresses may not yield representative traits of individuals. Additionally, many SDOH measures are based on sampling of all residents of a census division, but populations accessing health care may differ significantly from the rest of the individuals living in their neighborhood by virtue of needing health care. Address sources themselves can also serve as sources of spatial uncertainty. Patient addresses may simply be incorrect. This can be due to inaccurate transcription, ambiguous addresses, or out-of-date address records.[56] Finally, while our software package and scripts do not include non-U.S. geographic boundaries, if GIS data are available, they can be imported programmatically into our tool.

Conclusion

We developed an interactive, offline, web-based application to support address geocoding and mapping geocodes to neighborhood-level variables. POINT offers a HIPAA-compliant approach that can be easily scaled to multiple users with minimal technical expertise on a single installation. POINT successfully geocoded a greater percentage of addresses than existing geocoding tools. Among addresses that were successfully geocoded, we noted concordant mappings between systems which suggests accuracy. As health systems and researchers continue to explore and improve health equity, it is essential to obtain, and moreover, integrate into the EHR, accurate neighborhood level variables in a HIPAA-compliant way.

Clinical Relevance Statement

POINT is an offline geocoding solution that can support multiple users and integrates downstream mapping to neighborhood-level variables with a pipeline that allows users to incorporate additional datasets as they become available while protecting patient privacy. Geocoding at the patient level can enable targeted interventions that account for individual patient needs and circumstances based on the communities in which they live.

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Multiple-Choice Questions

1. What does it mean to geocode an address?
   a. Rewrite an address in a standardized form
   b. Convert the address into precise geographic coordinates
   c. Transfer an address into an electronic database
   d. Plot an address on a map

   Correct Answer: The correct answer is option b. Geocoding refers to the process of converting addresses from a text format (consisting of street number, street name, city, ZIP code, and state) into precise geographic coordinates (such as longitude and latitude).

2. What are community vital signs?
   a. Average heart rate, blood pressure, temperature, and respiratory rate of members of a given community
   b. Individual patient factors such as income or occupation
   c. Aggregate measures of SDOH in a community
   d. Average distance from a health care facility

   Correct Answer: The correct answer is option c. Community vital signs are measures of SDOH derived from neighborhood-level data. Like traditional vital signs, community vital signs provide clinicians with key information about the social environment in which patients live.

Protection of Human and Animal Subjects

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects and was reviewed by VUMC Institutional Review Board.

Funding

This work was funded by the National Institute on Aging (grant no.: R01AG062499).

Conflict of Interest

None declared.

References

1 Braveman P, Egerter S, Williams DR. The social determinants of health: coming of age. Annu Rev Public Health 2011;32(01): 381–398
2 Marmot M, Allen JJ. Social determinants of health equity. Am J Public Health 2014;104(suppl 4):S517–S519
3 Irwin A, Scali E. Action on the social determinants of health: a historical perspective. Glob Public Health 2007;2(03):235–256
4 Adler NE, Glymour MM, Fielding J. Addressing social determinants of health and health inequalities. JAMA 2016;316(16):1641–1642
5 Palmer RC, Ismond D, Rodriquez EJ, Kaufman JS. Social determinants of health: future directions for health disparities research. Am J Public Health 2019;109(S1):S70–S71
6 Davidson J, Vashisht R, Butte AJ. From genes to geography, from cells to community, from biomolecules to behaviors: the importance of social determinants of health. Biomolecules 2022;12(10):1449
7 Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129 (suppl 2):19–31
8 Norton JM, Moxey-Mims MM, Eggers PW, et al. Social determinants of racial disparities in CKD. J Am Soc Nephrol 2016;27(09): 2576–2595
9 Coughlin SS. Social determinants of breast cancer risk, stage, and survival. Breast Cancer Res Treat 2019;177(03):537–548
10 Avendano M, Glymour MM. Stroke disparities in older Americans: is wealth a more powerful indicator of risk than income and education? Stroke 2008;39(05):1533–1540
11 Gillum RF, Ingram DD. Relation between residence in the southeast region of the United States and stroke incidence. The NHANES I epidemiologic followup study. Am J Epidemiol 1996;144(07): 665–673
12 Reshetnyak E, Ntamatungiro M, Pinheiro LC, et al. Impact of multiple social determinants of health on incident stroke. Stroke 2020;51(08):2445–2453
13 Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care 2020;44 (01):258–279
14 Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc 2021;29(01):187–196
15 Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. Health Aff (Millwood) 2018; 37(04):585–590
16 Hatef E, Weiner JP, Kharrazi H. A public health perspective on using electronic health records to address social determinants of health: the potential for a national system of local community health records in the United States. Int J Med Inform 2019;124:86–89
17 Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. J Am Med Inform Assoc 2021; 28(12):2608–2616
18 Bazemore AW, Cottrell EK, Gold R, et al. "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. J Am Med Inform Assoc 2016;23(02):407–412
19 Center for Disease Control, Agency for Toxic Substances and Disease Registry. CDC/ATSDR Social Vulnerability Index data and documentation download. 2022. Accessed January 2, 2023 at: https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html
20 Centers for Disease Control. Environmental Justice Index (EJI). 2022. Accessed January 2, 2023 at: https://www.atsdr.cdc.gov/placeandhealth/eji/index.html
21 Agency for Healthcare Research and Quality. Social determinants of health database. Accessed January 2, 2023 at: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html
22 Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–the Public Health Disparities Geocoding Project. Am J Public Health 2002;92(07):1100–1102
23 Krieger N. A century of census tracts: health & the body politic (1906-2006). J Urban Health 2006;83(03):355–361
24 Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Record. Board on Population Health and Public Health Practice, Institute of Medicine.Capturing Social and Behavioral Domains and Measures in Electronic Health Records. National Academies Press (US); 2015. https://www.ncbi.nlm.nih.gov/books/NBK268995/
25 Brokamp C, Wolfe C, Lingren T, Harley J, Ryan P. Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. J Am Med Inform Assoc 2018;25(03):309–314
26 OpenStreetMap contributors. Accessed November 5, 2022 at: OpenStreetMap2017https://www.openstreetmap.org
27 Rashidian S, Dong X, Jain SK, Wang F. EaserGeocoder: integrative geocoding with machine learning (demo paper). In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. SIGSPATIAL '18. Association for Computing Machinery; 2018:572–575

 © 2023. Thieme. All rights reserved.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

28 PostGIS. PostGIS 3. 2022 Accessed November 5, 2022 at: https://postgis.net
29 Environmental Systems Research Institute. ESRI. ArcGIS Desktop
30 QGIS Association. QGIS geographic information system. Accessed November 5, 2022 at:http://www.qgis.org
31 Health Insurance Portability and Accountability Act of 1996. Pub. L. No. 104–191, § 264, 110 Stat.1936
32 US Census Bureau. US census TIGER/Line shapefiles file name definitions. Accessed November 5, 2022 at: www2.census.gov/geo/tiger/TIGER2022/2022_TL_Shapefiles_File_Name_Definitions.pdf
33 The PostgreSQL Global Development Group. PostgreSQL 15.1. 2022 Accessed November 5, 2022 at: https://www.postgresql.org
34 kevin-s-guo/point-geocoder. Accessed March 23, 2023 at: https://github.com/kevin-s-guo/point-geocoder
35 Uvicorn: ASGI web server implementation for python. Accessed November 5, 2022 at: https://github.com/encode/uvicorn
36 Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014(239):2
37 CDC Social Determinants of Health and PLACES Data. Accessed November 5, 2022 at: https://www.cdc.gov/places/social-determinants-of-health-and-places-data/index.html
38 US Department of Agriculture Economic Research Service. USDA ERS - Food Environment Atlas. Accessed January 2, 2023 at: https://www.ers.usda.gov/data-products/food-environment-atlas/
39 University of North Carolina at Chapel Hill. National health literacy mapping to inform health care policy. Health Literacy Data Map Accessed January 2, 2023 at: http://healthliteracymap.unc.edu/#
40 US Census Bureau. Community resilience estimates. Census.gov. Accessed January 2, 2023 at: https://www.census.gov/programs-surveys/community-resilience-estimates.html
41 Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. N Engl J Med 2018; 378(26):2456–2458
42 University of Wisconsin School of Medicine and Public Health. 2020 Area Deprivation Index. Accessed March 23, 2023 at: https://www.neighborhoodatlas.medicine.wisc.edu/
43 OpenAPI Initiative. OpenAPI Specification v3.1.0. Accessed November 5, 2022 at: https://spec.openapis.org/oas/v3.1.0
44 OpenAddresses. OpenAddresses: the free and open global address collection. Accessed January 5, 2023 at: http://openaddresses.io/
45 Harris DR, Delcher C. bench4gis: Benchmarking privacy-aware geocoding with open big data. In: 2019 IEEE International Conference on Big Data (Big Data); 2019:4067–4070. Doi: 10.1109/BigData47090.2019.9006234
46 Guo KS, Steitz BD. Open addresses nationally representative subset (1m). July 3, 2023. Doi: 10.5281/ZENODO.8112054
47 US Census Bureau. Batch address geocoder. 2022 Accessed November 5, 2022 at: https://geocoding.geo.census.gov/geocoder/geographies/addressbatch?form
48 Plotly Technologies Inc. Collaborative data science. 2015 Accessed November 5, 2022 at: https://plot.ly
49 Predmore Z, Hatef E, Weiner JP. integrating social and behavioral determinants of health into population health analytics: a conceptual framework and suggested road map. Popul Health Manag 2019;22(06):488–494
50 Hewner S, Casucci S, Sullivan S, et al. Integrating social determinants of health into primary care clinical and informational workflow during care transitions. EGEMS (Wash DC) 2017;5 (02):2
51 Ash AS, Mick EO, Ellis RP, Kiefe CI, Allison JJ, Clark MA. Social determinants of health in managed care payment formulas. JAMA Intern Med 2017;177(10):1424–1430
52 Lemke D, Mattauch V, Heidinger O, Hense HW. [Who hits the mark? A comparative study of the free geocoding services of Google and OpenStreetMap]. Gesundheitswesen 2015;77(8-9): e160–e165
53 Singh SK. Evaluating two freely available geocoding tools for geographical inconsistencies and geocoding errors. Open Geospatial Data Softw Stand 2017;2(01):11
54 Bell S, Wilson K, Shah TI, Gersher S, Elliott T. Investigating impacts of positional error on potential health care accessibility. Spat Spatio-Temporal Epidemiol 2012;3(01):17–29
55 Delmelle EM, Desjardins MR, Jung P, et al. Uncertainty in geospatial health: challenges and opportunities ahead. Ann Epidemiol 2022;65:15–30
56 Xierali IM. Physician multisite practicing: impact on access to care. J Am Board Fam Med 2018;31(02):260–269

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.