ORACLE

RWD 135

Curating Fit-for-Purpose Geographic Real-World Data

Authors: Dena H. Jaffe, PhD[1] and Amy Price, PhD[2]

Affiliations: 1. Oracle Health, Petah Tikva, Israel; 2. Oracle Health, Kansas City, MO, USA

Background

Greater geospatial granularity in real-world data (RWD) can improve measurement of local variation in healthcare quality, outcomes, and equity.

HIPAA de-identification methods of Safe Harbor and Expert Determination may necessitate reducing geographic precision. Under Safe Harbor, United States Postal Service (USPS) ZIP codes can be reported at the 3-digit prefix (ZIP3):[1]

(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(…)

(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:

(1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and

(2) The initial three digits of a zip code for all such geographic units containing 20,000 of fewer people is changed to 000."[1]

This de-identification process can, however, introduce information bias and obscure meaningful differences.

Objective

Describe privacy and data quality challenges when curating de-identified US geographic RWD using USPS ZIP codes and Census ZIP Code Tabulation Areas (ZCTAs) and provide practical curation recommendations for generating real-world evidence (RWE).

Methods

A comparative analysis of geographies was conducted using USPS ZIP codes and US Census Bureau data 5-digit ZCTAs. The first 3- and 5-digits of ZIP codes, ZIP3 and ZIP5, respectively were used as well as 3 and 5-digit ZCTAs, ZCTA3 and ZCTA5, respectively from the 2010 and 2020 censuses.[3-5] This study examined:

- HIPAA privacy implications under Safe Harbor and Expert Determination
- Data curation issues including evaluating ZIP5-ZCTA5 concordance using the Health Resources and Services Administration Data Warehouse (HRSA) crosswalk.[6,7]

Results

ZIP codes and ZCTAs are not equivalent

USPS ZIP codes and Census ZCTAs are related but distinct geographic units that are frequently conflated in RWE studies, with key differences in boundary definition, update frequency, and assignment logic (**Table 1**).

- ZIP codes are operational mail routing designations with no fixed geographic boundaries, and can change, split, or be discontinued as delivery patterns evolve
- ZCTAs are stable, bounded statistical approximations built by the Census Bureau by aggregating Census blocks, updated only at each decennial census
- Differences in the purpose of ZIP codes and ZCTAs should be considered for RWD curation, privacy, and linkage.

Table 1 Comparison of ZIP codes and ZCTAs

	USPS ZIP Code Used as **geographic locations** in healthcare data	**Census Bureau ZCTA** Used to measure **geographic characteristics**
Primary purpose	Mail delivery routing	Statistical analysis and data tabulation
Administering body	US Postal Service	US Census Bureau
Geometric form	No fixed areal boundary; defined by delivery routes (street segments, address ranges, and delivery points)	Areal units; polygon built from census blocks
Geographic coverage	Only areas receiving mail delivery	All populated/inhabited areas
Update frequency	Updated frequently (typically weekly); codes can be added, split, discontinued, or expanded at any time	Updated per decennial census (e.g., 2000, 2010, 2020)
Assignment logic	Based on mail delivery efficiency; may cross city, county, or state boundaries	Each census block is assigned to one ZCTA based on the most common ZIP code among addresses
Use in healthcare research	Common in administrative data (EHR, claims, surveys)	Preferred for spatial, demographic, and epidemiological analysis

HIPAA Privacy: Privacy rules can impact geographic resolution

Under HIPAA **Safe Harbor**, inhabited ZIP3 areas with Census populations below 20,000 must be recoded to 000 (**Table 2**).

The proportion of ZIP3 areas below this threshold increased from 1.3% (13 of 894) in 2010 to 2.0% (18 of 894) in 2020

Table 2 Inhabited areas in the United States with populations of <20,000 persons

ZIP3 codes that appear as ZCTA3	State	U.S. Census 2010	U.S. Census 2020
036	NH	13,759	13,153
059	VT	3525	3,352
102	NY	12,636	15,082
202	DC	N/A	0
203	DC	2055	772
204	DC	N/A	60
205	DC	N/A	44
369	AL	19,164	17,596
556	MN	16,024	16,415
692	NE	8637	8,626
753	TX	N/A	0
772	TX	N/A	4,280
821	WY	369	392
823	WY	16,430	14,702
878	NM	18,552	17,364
879	NM	17,432	16,082
884	NM	17,370	16,740
893	NV	12,103	10,701

Expert Determination treats geographies as quasi-identifiers. Certain ZIP codes associated with military, diplomatic, or federal government locations may warrant additional review because geographic detail can increase re-identification risk (**Table 3**).[10]

Table 3 ZIP3 Codes with quasi-identifying attributes

Federal Government	Military
005	090
055	091
192	092
202	093
203	094
204	095
205	096
375	097
399	098
459	099
509	340
649	962
733	963
842	964
938	965
	966

Inactive and special-use ZIP codes affect data quality

Non-inhabited, discontinued, and non-residential ZIPs can distort joins, denominators, and interpretation.

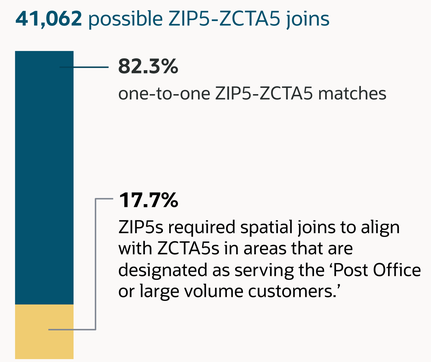
Flag, aggregate, or suppress ZIP3s that do not represent residential geography, for example, post office (PO) box/parcel center, institutional, and discontinued ZIP codes (**Table 4**). Treating these as residential or currently populated areas can create geospatial bias, such as misattribution of outcomes or inflated denominators.

Table 4 USPS ZIP3 codes to consider for improved data quality

Post office box/Parcel return centers	Non-government Organizations	Not in use/Other			
311	889	000	552	698	862
332	901	003	555	699	866
569		213	568	702	867
753		269	578	709	868
772		343	579	715	869
885		345	589	732	872
942		348	621	742	876
		353	632	771	886
		419	642	817	887
		428	643	818	888
		429	659	819	889
		517	663	839	892
		518	682	848	896
		519	694	849	909
		529	695	854	929
		533	696	858	987
		536	697	861	

To improve geographic accuracy, USPS ZIP5 codes should be cross-walked to 5-digit ZCTAs when feasible. Figure 1 illustrates the use of the HRSA crosswalk [6,7]:

Figure 1 Concordance of ZIP5 and ZCTA5



<0.01% Populated ZCTA5 & missing ZIP5
<0.01% Territory ZIP5 & no ZCTA5 available

Conclusion

Curation of geographic data is critical to ensuring the validity and relevance of RWD for RWE.

Privacy-compliant geographic curation considerations:

- Understanding the structural differences between USPS ZIP codes and Census Bureau ZCTAs
- Ongoing validation of ZIP codes to account for postal route and code changes and population shifts
- Identification of areas that may be indirect patient identifiers and may warrant HIPAA Expert Determination review.

Data quality geographic curation considerations:

- Suppression, aggregation, or flagging of PO box, institutional, and non-residential ZIP codes can introduce geospatial bias by attributing risk or outcomes to locations with no residential populations and inflating population denominators
- Adjustment for spatial mismatch between ZIP codes and ZCTA boundaries for population descriptors, can bias data linkage
- Identifying geographic discontinuities in longitudinal analyses can create artifactual breaks in disease tracking
- Recognition that aggregation of less populated areas can mask at-risk populations and health disparities.

Researchers should understand the limitations of USPS and Census geographies, document geographic transformations, and be aware of potential sources of bias.

Mr. ZIP 1963[2]

References:

1. U.S. Government Publishing Office. https://www.ecfr.gov/current/title-45/subtitle-A/subcha pter-C/part-164/subpart-E. Accessed 10 July 2025.; **2.** USPS, 2008 https://about.usps.com/who/profile/history/pdf/mr-zip.pdf; **3.** U.S. Census Bureau. 2023. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html; **4.** U.S. Census Bureau, 2026a. https://data.census.gov/table/DECENNIALSF12010.P1?q=All+5-digit+ZIP+Code+Tabulation+Areas+within+United+States. Accessed 12 Apr 2026.; **5.** U.S. Census Bureau, 2026b. https://data.census.gov/table/DECENNIALDHC2020.P1?q=All+5-digit+ZIP+Code+Tabulation+Areas+within+United+States. Accessed 12 Apr 2026.; **6.** Health Resources and Services Administration (HRSA). https://data.hrsa.gov/DataDownload/GeoCareNavigator/ZIP%20Code%20to%20ZCTA%20Crosswalk.xlsx. Accessed 4 Apr 2026; **7.** U.S. ZIP Code Points Feature Layer. https://www.arcgis.com/home/item.html?id=dc123f738bf846779c49db6472f82a4b. Accessed 4 Apr 2026; **8.** Grubesic & Matisziw. *Int J Health Geogr* 2006;5:58; **9.** Wilson & Din. *Cityscape* 2018;20(2):277–294; **10.** Xie et al. *AMIA Jt Summits Transl Sci Proc.* 2023;2023:572-581; **11.** Jaffe et al. *Health Affairs Scholar* 2025;3:qxaf210.